CSC3065 Cloud Computing
Assessment 3: Search Engine

Assessment Briefing Version:	1 (03/11/2019)
Weighting:	40%
Set By:	David Cutting
Moderated By:	Neil Anderson

Date Released:	**05/12/2019**
Submission Due:	**1700 on 07/01/2019**

Late submission penalties and rules will be applied in accordance with the QUB policy on late submission. For more information on this or any other QUB policy with regards to assessment please see:

https://www.qub.ac.uk/directorates/AcademicStudentAffairs/AcademicAffairs/Examinations andAssessment/MarkSchemesandClassifications/

Please be aware that this is an individual assignment and as such will be checked for plagiarism. You should ensure that the work is yours and yours alone, citing any third-party sources as applicable. Plagiarism is a serious academic offense. Submission of your work implies your claim that it is your individual work and has not previously been submitted for academic credit.

If you have any questions about the assignment please see the module organiser David Cutting in the first instance or use any of the support options listed in this document.

1. Assessment Details

The Queen's Search Engine (QSE) will be an innovative and amazing search platform using fantastic new techniques and technologies never seen before, apart from in pretty much any other search system.

Your task: architect, implement, deploy, manage, and monitor a search engine system. Through the submission (section 4) you should clearly show all aspects of your system. For the architecture especially this is expected to contain diagram(s), decomposed as appropriate to different levels of the system, and any justifications or design decisions made.

Please note that unlike previous tasks here there is no penalty for using pre-made third party components, in fact to accomplish QSE in a timely manner it's probably going to be essential.

Please also note that the challenge here is not about any of the algorithmic challenges around indexing and searching (which is a whole field in itself) but focussed specifically on the cloud architecture and operation of QSE. QSE should be highly scalable and very resilient.

The most highly marked projects will be highly flexible and scalable solutions demonstrating multi-provider management and technologies such as pipelines, monitoring, deployment, etc as appropriate with clearly justified decisions.

1.1 Primary Components

QSE consists of three main components which interact with the outside world:

Spider Indexer – something that can crawl the web and collect data from pages. If this works with live pages it should be a "good citizen" and implement best practice.

Note there's an interesting starter guide here which may provide some overall inspiration for web indexing and has a good set of best practices: https://medium.com/velotio-perspectives/web-scraping-introduction-best-practices-caveats-9cbf4acc8d0f

The indexer will then put the found data into some sort of stored QSE index.

Search – querying the QSE index against specific search criteria. Again, remember this is not a challenge about most effective search algorithms, pretty standard searches are absolutely fine, it's about how you architect your system to allow simultaneous search at scale.

Ads – a directory of adverts with associated keywords which could be displayed alongside search results or theoretically embedded onto third-party sites and services.

To support this there will need to be one, or more, internal data store(s) of indexed data.

In addition to the external components and the data store(s) QSE also ideally requires:

Deployment and Management – some system to deploy and manage the QSE components.

Monitoring – a system which monitors the health of all the components and services of the QSE which ideally would include alerting.

Along with any other components or features which make operation or management of the system easier/more efficient.

1.2 Optional Provided Resources

To get you started and/or act as an inspiration and/or be used in the final solution as a placeholder if some aspects of your QSE don't work with real data some <u>very basic</u> resources are provided.

Note: if you do include these into your final submission you must clearly say so. For full marks you must generate your own data and use your own components, but use of these will allow demonstration of functionality.

https://files.davecutting.uk/CSC3065/Assessment3-Resources.zip

Contained in this zip archive are:

- In the wiki folder several Wikipedia pages in pure text format (note the effective link of each of these is en.wikipedia.org/wiki/X where X is the case-sensitive filename)
- In the index folder a very basic SQL schema and data set containing the pages from the wiki folder – these are exported from MySQL but should be generic enough to run in most database engines
- In the ads folder a very basic SQL schema and data set containing some simple adverts

Note that these in no way are intended to represent any sort of suitable data structure or format, just a placeholder to get started with!

1.3 Platform

As with everything else this time the choice of platform(s) is entirely up to you. Over the Christmas break accessing the QPC may be a hassle so it may make sense to look at other providers using the educational accounts provided.

Take note of the top end descriptors for a number of the marking criteria which require the demonstration of multiple provider management and orchestration (building a highly resilient system beyond a single provider).

2. Assessment Criteria

The following are the criteria against which your submission will be marked and their conceptual marking equivalents.

Marking Criteria

Criteria	Outstanding 85% +	Excellent 70%-85%	Very Good 60%-70%	Good 50%-60%	Acceptable 40%-50%	Unacceptable < 40%
Architectural Design (30%)	An outstanding design that clearly shows a well deconstructed architecture which offers great resilience and scalability over multiple providers. Clear and suitable presentation of an outstanding level.	An excellent design that clearly shows a well deconstructed architecture which offers great resilience and scalability over multiple providers with some room for improvement. Excellent presentation.	A very good design and shows a strong architecture offering high levels of resilience and scalability over one or more providers with areas of improvement. Well and clearly presented.	A good design suitable for some good levels of scalability and resilience but with significant room for improvement. Presentation of a generally good standard.	A rudimentary design offering some scalability and resilience only. Presentation may have significant issues in presentation and clarity.	An unsuitable or lacking design with little or no scalability or resilience. Weaknesses in presentation style and clarity.
Implementation of each external component (index, search, ads) (3 x 10%)	Excellent implementation of each service with industry-ready approaches and techniques used as appropriate.	Very good implementation of each service generally using industry-ready approaches but with a few minor areas for improvement.	Good implementation of each service generally using best practices with areas for improvement in process or function.	Services implemented and functional but lacking in some respects of the approach, quite a few areas of improvement.	Services implemented but with shortcomings in function or process with significant areas of potential improvement.	Services very poorly partially or not implemented

Deployment scripting and management (20%)	Outstanding highly flexible provider-agnostic deployment orchestration and management scripts to an industry-ready standard for live services. Working with multiple providers as needed.	Excellent highly flexible provider-agnostic deployment orchestration and management scripts to a near industry-ready standard for live services. Working with multiple providers as needed. Some minor areas which could be improved.	Very good generally flexible provider-agnostic deployment orchestration and management scripts (or provider-specific but flexible) to a high standard. Some areas which could be improved.	Good generally flexible deployment orchestration and management scripts/controls to a good standard. Grounds for significant improvement.	Some generally flexible deployment orchestration and management controls demonstrated working to an acceptable standard. Numerous weaknesses in approach and implementation.	Little to no evidence of appropriate use of deployment and management tools.

Monitoring and reporting (20%)	Outstanding monitoring and reporting demonstrating industry-ready approaches to working with live services. All functionality including alerts and historic reporting working completely.	Excellent monitoring and reporting generally demonstrating industry-ready approaches to working with live services. Nearly all functionality including alerts and historic reporting working completely.	Very good monitoring and reporting generally demonstrating best practice approaches to working with live services. Some areas for improvement.	Good monitoring and reporting generally demonstrating strong and applicable approaches to working with live services. Some significant areas for improvement.	Acceptable monitoring and reporting with most key services monitored. Significant areas for improvements.	Mostly missing or faulty monitoring with little to no reporting.

3. Feedback

Feedback in the form of marks will be provided as soon as practicable after submission with the expectation that marking will be complete (and marks provided) within two working weeks.

Individual feedback will take the form of a numeric score against each of the assessment criteria (there may be brief comments for these criteria if appropriate), a total made from these scores weighted by section, and an overall textual comment on the totality of the submission.

Generalised feedback will be provided to the class as a whole including overall trends and areas of particular concern.

Anyone wishing to discuss their marks in more detail are welcome to do so using any of the support arrangements outlined in this document.

4. Submission

Submission will be via Canvas and also through repository access. Please read this section carefully to avoid any mistakes which could lead to marks being lost.

You are required to:

1. Provide repository access if not using the QPC gitlab (gitlab.hal.davecutting.uk – if you are using this then you don't need to do anything other than provide the links in your report) so for example the EEECS gitlab or any other repository service. It is your responsibility to discuss this with a member of the teaching team to allow them access (or a link to a public repository) if this is the case.

2. Upload a ZIP the filename of which contains your student number, the ZIP file must contain:
- A completed report based on the template provided
- A demonstration video showing the operation of your system (see 4.1 for more details)
- A copy of all source code you have written (this is required in addition to the repository access for the external examiner)

4.1 Video Submission

To demonstrate your system working you are required to submit a video, usually a screen capture video, ideally as short as possible. This should show each of the key elements of your solution and may include reference to the architecture diagrams provided.

You should use this video to highlight any specifics you wish to, for example any particularly detailed functionality included.

5. Support Available

A number of support avenues are available throughout this project. It's suggested you try them in this order, but this is your choice and you should feel free to avail yourself of one or all.

Canvas Discussion – you can ask any questions (in general please, don't include your work as everyone can see!) on the Canvas Discussion Forum. This is very useful as everyone can see (the question and the answer!) and its possible students can help each other out.

Module Drop-in – A specific drop-in for this assignment will be held in week 12 on Wednesday 11th December 1100-1300.

Online Support Seminars – Using some technology (most likely Canvas conferences) there will be the following online sessions you are all welcome to join (and ask questions in). Please note that owing to the Xmas break getting individual email answers may be significantly delayed so I strongly urge you to virtually attend these sessions with any questions you may have. They will start at the given time and continue until questions have been answered (so if you're late they may have finished!). You can pre-email questions to get answered or ask live in the session.

Specific technical details on how to join will be shared via announcements nearer the time.

Tuesday 17th December   1300
Monday 23rd December   1500
Monday 30th December   1500

Open Door/Other Appointments – David Cutting operates an "open door" policy when he is available but it's advisable to fire off an email first to check he's available. He is happy to make ad-hoc meetings for individuals or groups as needed outside of the office hours above. Please note that he is unlikely to be regularly in the office over the Xmas period, if necessary Skype appointments can be arranged but primarily focus will be on the Online Support Seminars mentioned above.