Informatics Board of Studies - Course Proposal

Proposed course title: Data Architectures (by Distance Education)

Proposer(s): Gary McGilvary, Maria Wolters, Michelle Galea, Dave Robertson

Date: 5/01/15

1. Case for Support

[This section should summarise why the new course is needed, how it fits with the existing course portfolio, the curricula of our Degree Programmes, and delivery of teaching for the different years it would affect.]

1a. Overall contribution to teaching portfolio

[Explain what motivates the course proposal, e.g. an emergent or maturing research area, a previous course having become outdated or inappropriate in other ways, novel research activity or newly acquired expertise in the School, offerings of our competitors.]

The School of Informatics wishes to expand and broaden its reach by providing quality educational experiences to students who may not be able or may not wish to take on-campus courses in Edinburgh. Therefore the School of Informatics has been establishing a number of initiatives to support the creation of online education offerings for the last few years and now seeks to create initial courses and modules destined for fee earning online and distance education uses.

The Data Science effort, lead by Dave Robertson, aims to build a modular MSc in Data Science distance education programme; an area identified as being within the top four market opportunities for distance education in Informatics 2012. Data science is the study of the computational principles, methods, and systems for extracting knowledge from data. Large data sets are now generated by almost every activity in science, society, and commerce – ranging from molecular biology to social media, from sustainable energy to health care.

Data science asks: how can we efficiently find patterns in these vast streams of data? Many research areas have tackled parts of this problem: machine learning focuses on finding patterns and making predictions from data; ideas from algorithms and databases are required to build systems that scale to big data streams; and separate research areas have grown around different types of unstructured data such as text, images, sensor data, video, and speech. Recently, these distinct disciplines have begun to converge into a single field called Data Science.

We aim to develop and provide a masters-level programme focussing on the theme of Data Science, by making an early offering of relevant modules via the promotion of an interim PGCertificate (60 points) in this area. Our initial plan sees students taking an initial 20 point core course introducing the basics of Data Science and further 10 point elective modules in a specialism they choose.

One such elective module, which we propose here is the Data Architectures course.

The Data Architectures course will provide an overview of the principles, infrastructures and algorithms behind analysing large-scale data analysis and problem solving. It will touch upon the technologies and techniques used by companies such as Google, Amazon, Facebook and Microsoft to enable warehouse-scale computing. The course will be in three parts: the principles behind Data Architecture infrastructures (cloud computing, scaling, performance, privacy, etc), supporting infrastructures (distributed file systems, replication, web services, etc) and algorithms (MapReduce, case studies from Natural Language Processing, Database query evaluation, machine learning, streaming).
As the current UG4 and MSc course Extreme Computing is particularly relevant to the architectural aspects of Data Science, we will include much of the Extreme Computing lecture and practical material into our Data Architectures course. Therefore the method of delivery for this course is the only substantial difference.

This distance learning offering will allow interested students to study in-demand topics without the constraints of campus attendance, and School staff to develop skills and experience in the development and delivery of distance education programmes, and in the support of distance education students. It will also provide the School with another opportunity to establish new relationships across industry, the public sector and academia.

While many other universities offer online courses similar to our Data Architectures course, either at MSc or MOOC level, the success of our Data Architectures course will primarily rely on the uptake of our MSc/PGCertificate in Data Science by distance education programme, which will be driven by the fact that students’ criteria for choosing a university will be driven by the quality of the institution, the educational experience and social media communications about that to a greater extent.

1b. Target audience and expected demand

Describe the type of student the course would appeal to in terms of background, level of ability, and interests, and the expected class size for the course based on anticipated demand. A good justification would include some evidence, e.g. by referring to projects in an area, class sizes in similar courses, employer demand for the skills taught in the course, etc.

Students are expected to be primarily professionals in data-intensive industries looking to extend their expertise.

Background: we would expect a Computer Science related background and have the ability to program.

Demand: while our Data Science PGCertificate (by Distance Education) will be offered to the general public, we will primarily market the individual courses, such as this Data Architectures course, and the PGCertificate offerings to the wide range of small and large businesses and government agencies we collaborate and interact with. For example, through our various initiatives, commercialisation unit, the Data Science centre for doctoral training, and relevant innovation centres.

Market research in early 2014 examined industry activities relating to Data Science. A summary of the survey results relevant to distance education in Data Science are outlined below:

- 69 responses from a number of public and private sector organisations. The primary focus was on identifying industry demand.
- 45% of respondents would look to engage with a university to further knowledge/skills.
- Companies spanned a number of industries: Aerospace, Defence, Chemicals, Creative Industry, Education, Engineering, Energy, Financial Services and others.
- Only 1% of respondents had previously worked with a Scottish university through CPD.
- But 63% of respondents either said they were Somewhat Interested, Interested or Very Interested in such Continued Professional Development/Training in the future.

We will also apply for funding to offer scholarships and fee-support to distance education students to increase the range of people who may be able to benefit, from industry or not. We anticipate many students taking individual modules and particularly the Data Architectures course, as the infrastructure, technologies and techniques taught are extremely relevant for those working in industry and for those who wish to increase their employability in this area.
1c. Relation to existing curriculum

This section should describe how the proposed course relates to existing courses, programmes, years of study, and specialisms. Every new course should make an important contribution to the delivery of our Degree Programmes, which are described at http://www.drps.ed.ac.uk/12-13/dpt/drps_inf.html. Please name the Programmes the course will contribute to, and justify its contribution in relation to courses already available within those programmes. For courses available to MSc students, describe which specialism(s) the course should be listed under (see http://www.inf.ed.ac.uk/student-services/teaching-organisation/taught-course-information/year-guides/taught-postgraduate-year-guide/degree-requirements/specialist-areas), and what its significance for the specialism would be. Comment on the fit of the proposed course with the structure of academic years for which it should be offered. This is described in the Year Guides linked from http://www.inf.ed.ac.uk/student-services/teaching-organisation/taught-course-information/year-guides/.

As previously mentioned, the proposed Data Architectures course will be synonymous to the current UG4 and MSc Extreme Computing course. We expect to inherit much of the lecture and practical material and deliver this material through various distance education platforms.

The Data Architectures course relates to the following MSc specialisms: Analytical & Scientific Databases, Computer Systems, Software Engineering & High Performance Computing, and Learning from Data.

1d. Resources

While course approvals do not anticipate the School's decision that a course will actually be taught in any given year, it is important to describe what resources would be required if it were run. Please describe how much lecturing, tutoring, exam preparation and marking effort will be required in steady state, and any additional resources that will be required to set the course up for the first time. Please make sure that you provide estimates relative to class size if there are natural limits to its scalability (e.g. due to equipment or space requirements). Describe the profile of the course team, including lecturer, tutors, markers, and their required background. Where possible, identify a set of specific lecturers who have confirmed that they would either like to teach this course apart from the proposer, or who could teach the course in principle. It is useful to include ideas and suggestions for potential teaching duty reallocation (e.g. through course sharing, discontinuation of an existing course, voluntary teaching over and above normal teaching duties) to be taken into account when resourcing decisions are made.

Below are our estimations of effort required by the course team for the following tasks:

- **Course Preparation:** The Extreme Computing course currently has relevant lecture (not video) and practical material, which could be re-used and delivered without major changes. Much of the effort will be spent integrating the current practical material with a platform we have chosen to offer a practical development environment for students to work in. This platform is supplied by Aridhia and they have agreed work as part of the course production team to produce practical exercises that can be run on-line via their analytics toolkit. University IS have agreed to support the toolkit on its servers and are dealing with contractual arrangements.

  Currently, the Extreme Computing labs and practicals are based on Hadoop and we aim to use the Hadoop console of the Aridhia platform to run these practicals. As an upper estimate, 3-4 person weeks may be required to convert existing material, create new material as necessary and integrate material with the various distance learning platforms we have chosen, but funding for that already exists and will not impact on Informatics staff.

- **Lectures:** minimal as material will be made available online from Moodle hosted by the University and the student will make their way through the material.

- **Lab sessions:** The equivalent of lab sessions (for a traditional, residential course) will be exercises through which the student will make their way through (supported by access to a tutor
on-line). We expect these to be run by current or new tutors, mainly current PhD students, especially in the first few years when student numbers will be at their lowest.

- **Coursework Marking effort:** we aim for practical assignments and exams to be marked by members of the course team, or via PhD students, especially in the first few years when student numbers may be at their lowest.

- **Student interaction:** it is important that distance learning students feel a part of both the School and University. Resources are therefore required to engage with and encourage them to work closely with the course team, and to link to peer-to-peer support via the Moodle and School social platforms, discussion forums, and specific course/session levels. One course tutor would be required to answer student questions, mark or manage marking of practical assignments, provide feedback, etc.

- **Lab/Coursework Resources:** As a distance education offering, course lab/tutorial work and practical assignments will make use of a University hosted platform, provided by Aridhia, to allow students to program and run code without the need to download software locally, e.g. R or Hadoop.

- **Exam preparation:** Minimal as exams will be prepared either in collaboration with current Extreme Computing organizers or from previous Extreme Computing exams.

- **Course team:** As part of the wider Data Science effort, a core team currently exists to support various aspects related to all distance education courses, e.g. manage underlying distance education platforms, provide administration, etc. One course tutor would be required to answer student questions, mark or manage marking of practical assignments, provide feedback, etc. In the event that an extremely large number of students sign up, either in the first year or in the future, more staff would be required to support students, mark work, prepare exams, etc. This can be funded by the income generated from these signups.
2. Course descriptor

[This is the official course descriptor that will be published by the University and serves as the authoritative source of information about the course for students. Current course descriptions in the EUCLID Course Catalogue are available from http://www.star.euclid.ed.ac.uk/ipp/cx_sb_infr.htm.]

Course Title: Data Architectures

SCQF Credit Points: 10
SCQF Credit Level: 11
Normal Year Taken: MSc
Also available in years: NA

Subject Area and Specialism Classification: Computer Science, Software Engineering, Analytical & Scientific Databases, Computer Systems, Software Engineering & High Performance Computing, and Learning from Data

Appropriate/Important for the Following Degree Programmes: MSc/PGCertificate in Data Science by Distance Education

Timetabling Information: Semester 1 and 2 (no timetable set)
To be determined as part of our distance education marketing.

School Acronym: INF-???

Short Course Description: Data Architectures deals with the principles, systems and algorithms behind Web-scale problem solving. This touches upon the technologies and techniques used by companies such as Google, FaceBook and Microsoft, using warehouse-scale computing and massive datasets. The course will be in three parts: the principles behind Data Architectures (cloud computing, scaling, performance, privacy etc), supporting infrastructure (distributed file systems, replication, web services etc) and algorithms (MapReduce, case studies from Natural Language Processing, Database query evaluation, machine learning, streaming).

Pre-Requisite Courses: None

Co-Requisite Courses: None

Prohibited Combinations: None

Other Requirements:
A Maths background is required, including basic probability as well as programming ability and familiarity with Unix-like systems. Any programming language is fine; past students find that Python is sufficient.

Available to Visiting Students: No
Summary of Intended Learning Outcomes:

1. Be able to identify current issues and trends in the area of Data Architectures.
2. Gain experience in applying data science methods in practice.
3. Demonstrate knowledge of the need for large-scale data analysis by providing motivating examples of the scale of problems only large-scale data analysis can solve.
4. Demonstrate knowledge of the infrastructure necessary for Data Science through enumerating different file system designs, virtualisation techniques, and fault-tolerance paradigms.
5. Demonstrate knowledge of cluster-based algorithms for natural language processing, database query evaluation, machine learning, and data stream processing through the use of the Map/Reduce programming paradigm.

Assessment Information

Students should expect to spend approximately 40 hours on the coursework for this course. The coursework is composed of two practical assignments, both of which test the students’ ability to develop and experiments with algorithms on a Hadoop/MapReduce infrastructure. Assessment may also include use of online questionnaires and progress quizzes, peer assessment, and class participation.

There will be two “lab” sessions available before each coursework assignment for students to get familiarised with the infrastructure and the concepts they will have to use for the practical assignments.

As this course is offered as a distance learning course, we envisage that students will partake in an online and timed exercise at the end of the course.

The above learning outcomes are tested both throughout the assessed coursework and written examination.

Assessment Weightings:

- Written Examination: 70%
- Practical Examination: 0%
- Coursework: 30%

Time spend on assignments: 40 hours

Syllabus:
The course will conceptually be split into three main areas, with each area not necessarily accounting for an equal portion of the syllabus. The three areas and the material covered in each area are as follows:

- Background: Motivation for new computing paradigms; introduction and differences between cloud and cluster computing; scaling, performance, privacy, economics, security, software as service.
• Infrastructure: Distributed file systems; virtualisation; replication; fault tolerance; concurrent programming; web services.

• Algorithms: Hadoop (MapReduce); design and implementation of MapReduce programs; dealing with massive amounts of data; case studies using natural language processing, database query evaluation and machine learning; data stream processing.

**Relevant QAA Computing Curriculum Sections:** Computer Based Systems, Computer Networks, Concurrency and Parallelism, Databases, Data Structures and Algorithms, Developing Technologies, Distributed Computer Systems, Information Retrieval, Natural Language Computing

**Reading List:**
- Data Intensive Text Processing with MapReduce, Jimmy Linn & Chris Dyer
- Hadoop: The Definitive Guide, Tom White, O'Reilly Media

**Breakdown of Learning and Teaching Activities:**

20 lecture hours and 0 tutorial hours each week, with 2 coursework assignments and 2 lab practicals.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture Hours</td>
<td>20</td>
</tr>
<tr>
<td>Seminar/Tutorial Hours</td>
<td>0</td>
</tr>
<tr>
<td>Supervised practical/Workshop/Studio hours</td>
<td>0</td>
</tr>
<tr>
<td>Summative assessment hours</td>
<td>40</td>
</tr>
<tr>
<td>Feedback/Feedforward hours</td>
<td>5</td>
</tr>
<tr>
<td>Directed Learning and Independent Learning</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>
3. Course materials

3a. Sample exam question(s)
[Sample exam questions with model answers to the individual questions should be provided. A justification of the exam format should be provided where the suggested format non-standard. The online list of past exam papers gives an idea of what exam formats are most commonly used and which alternative formats have been http://www.inf.ed.ac.uk/teaching/exam_papers/]

Previous Extreme Computing exam papers will be similar to Data Architectures exams:
https://exampapers.ed.ac.uk/record/48185/1/20141224.pdf

3b. Sample coursework specification
[Provide a description of a possible assignment with an estimate of effort against each sub-task and a description of marking criteria.]

Previous Extreme Computing coursework will be similar to Data Architectures coursework:
http://www.inf.ed.ac.uk/teaching/courses/exc/assignment.html

3c. Sample tutorial/lab sheet questions
[Provide a list of tutorial questions and answers and/or samples of lab sheets.]

Previous Extreme Computing lab assignments will be similar to Data Architectures lab assignments:
http://www.inf.ed.ac.uk/teaching/courses/exc/labs.html

4. Course management

4a. Course information and publicity
[Describe what information will be provided at the start of the academic year in which format, how and where the course will be advertised, what materials will be made available online and when they will be finalised. Please note that University and School policies require that all course information is available at the start of the academic year including all teaching materials and lecture slides.]

Course content will be served from Moodle hosted by the University, initially hosting the lecture set and reading list. The current Extreme Computing lecture set can be used, perhaps subject to minor modifications. Similarly, lab/tutorial work and practical assignments can be taken from the Extreme Computing course, and these can be added as the student progresses.

As a distance education offering, course lab/tutorial work and practical assignments will make use of a University hosted platform, provided by Aridhia, to allow students to program and run code without the need to download software locally, e.g. R or Hadoop.

This course will be advertised alongside our marketing of the PGCertificate in Data Science programme, which will include Informatics website advertising, making relevant industries aware, etc. We have dedicated administrative support for this in the first year, funded by the Distance Education Initiative.
4b. Feedback
Provide details on feedback arrangements for the course. This includes when and how course feedback is solicited from the class and responded to, what feedback will be provided on assessment (coursework and exams) within what timeframe, and what opportunities students will be given to respond to feedback. The University is committed to a baseline of principles regarding feedback that we have to implement at every level, these are described at http://www.docs.sasg.ed.ac.uk/AcademicServices/Policies/Feedback_Standards_Guiding_Principles.pdf. Further guidance is available from http://www.enhancingfeedback.ed.ac.uk/staff.html.

We plan a rich and well-resourced level of engagement between distance education students and world-leading Informatics teaching and research staff:
• Extensive use of the School and University level virtual learning environments (VLE) and social platform is planned.
• Course forums will allow students to ask questions to both teaching staff and to other students.
• Virtual class presentation and project spaces will be available based on the extensive experience of the School with the Virtual University of Edinburgh (Vue) facilities.

The course tutor will comment on each student’s work either via Moodle or a VLE.

4c. Management of teaching delivery
Provide details on responsibilities of each course staff member, how the lecturer will recruit, train, and supervise other course staff, what forms of communication with the class will be used, how required equipment will be procured and maintained. Include information about what support will be required for this from other parties, e.g. colleagues or the Informatics Teaching Organisation.

We expect that the course tutor simply needs to provide support to course students and flag any issues that do arise related to the delivery of the course. As this course is part of the Data Science Distance Education effort, the Data Science team or University support teams will handle most issues concerned with content delivery, e.g. issues with the Aridhia platform or Moodle, respectively.

Communication with the class will primarily occur via Moodle and a course mailing list however at times, the University level VLE’s will also be used; all of which are supported by the relevant teams. Minimal support is required from the ITO.

An important part of teaching delivery is the engagement with the students. As these students will become formally enrolled at the university, we want to create an environment where they feel they part of the university and school despite the distance. Firstly, we need platforms such as Moodle to engage and encourage students, allow peer-to-peer support, engage in discussions etc. And secondly, we need the staff resources (estimated as 1 tutor for the first year) to create and maintain this environment.
5. Comments

[This section summarises comments received from relevant individuals prior to proposing the course.]

This is one of the first of a new line of distance education courses to be run within a cross-College framework so the concepts of year organiser and degree programme coordinator (though still applicable) are not the same as for our residential courses. We are in discussions with our Assistant Principal Learning and Development to build an appropriate “container” for these courses with associated academic and administrative responsibilities.

5a. Year Organiser Comments

[Year Organisers are responsible for maintaining the official Year Guides for every year of study, which, among other things, provide guidance on available course choices and specialist areas. The Year Organisers of all years for which the course will be offered should be consulted on the appropriateness and relevance on the course. Issues to consider here include balance of course offerings across semesters, subject areas, and credit levels, timetabling implications, fit into the administrative structures used in delivering that year.]

5b. Degree Programme Co-Ordinators

[Degree Programme Co-Ordinators are responsible for maintaining the official Degree Programme Specifications and Degree Programme Table for a given subject area which, among other things, specify the content of courses taken in a Degree Programme. The Degree Programme Co-Ordinators of the relevant subject areas that the course is proposed for should comment on the fit with the current curriculum of the relevant Degree Programmes. Issues to consider here are dependencies arising from pre-, co-requisites, and forbidden combinations, balance of different topics in a Degree Programme, etc.]

5c. BoS Academic Secretary

[Any proposal has to be checked by the Secretary of the Board of Studies prior to discussion at the actual Board meeting. This is a placeholder for their comments, mainly on the formal quality of the content provided above.]

Comments received from a number of staff; these are addressed in the text above.