SQA Advanced Unit specification

General information

Unit title: Data Science (SCQF level 8)

Unit code: HR9V 48

Superclass: CB

Publication date: August 2017

Source: Scottish Qualifications Authority

Version: 01

Unit purpose

The purpose of this Unit is to provide an introduction to the theory and practice of data science. The Unit is for those who already have some knowledge of big data and wish a deeper understanding of the techniques used to process it. It is suitable for a wide range of learners.

The Unit covers a mix of theory and practice. The theoretical content includes the concepts behind data science, the techniques used to store and manage big data and the tools used to process it. The practical content relates carrying out data analysis.

The Unit seeks to build on existing knowledge of this emerging discipline so that learners can appreciate its actual and potential uses in a range of contexts.

Outcomes

On successful completion of the Unit the learner will be able to:

1 Describe the main areas of operation of data scientists.
2 Describe the techniques used to store and manage big data.
3 Describe the principal tools used to manipulate big data.
4 Use tools to carry out data analysis.

Credit points and level

2 SQA Credits at SCQF level 8: (16 SCQF credit points at SCQF level 8)

Recommended entry to the Unit

It would be beneficial is learners had completed the SQA Advanced Unit *Big Data* at SCQF level 7 and possessed *Numeracy Skills*. This may be evidenced by possession of the Core Skills Unit in *Numeracy* at SCQF level 6. Some previous knowledge of statistics is desirable but not essential.

Core Skills

Opportunities to develop aspects of Core Skills are highlighted in the Support Notes for this Unit specification.

There is no automatic certification of Core Skills or Core Skill components in this Unit.

Context for delivery

If this Unit is delivered as part of a Group Award, it is recommended that it should be taught and assessed within the subject area of the Group Award to which it contributes.

The Assessment Support Pack (ASP) for this Unit provides assessment and marking guidelines that exemplify the national standard for achievement. It is a valid, reliable and practicable assessment. Centres wishing to develop their own assessments should refer to the ASP to ensure a comparable standard. A list of existing ASPs is available to download from SQA's website **(http://www.sqa.org.uk/sqa/46233.2769.html)**.

Equality and inclusion

This Unit specification has been designed to ensure that there are no unnecessary barriers to learning or assessment. The individual needs of learners should be taken into account when planning learning experiences, selecting assessment methods or considering alternative evidence.

Further advice can be found on our website **www.sqa.org.uk/assessmentarrangements**.

Unit specification: Statement of standards

Unit title: Data Science (SCQF level 8)

Acceptable performance in this Unit will be the satisfactory achievement of the standards set out in this part of the Unit specification. All sections of the statement of standards are mandatory and cannot be altered without reference to SQA.

Where evidence for Outcomes is assessed on a sample basis, the whole of the content listed in the Knowledge and/or Skills section must be taught and available for assessment. Learners should not know in advance the items on which they will be assessed and different items should be sampled on each assessment occasion.

Outcome 1

Describe the main areas of operation of data scientists.

Knowledge and skills

- Data architecture
- Data acquisition
- Data analysis
- Data archiving

Outcome 2

Describe the techniques used to store and manage big data.

Knowledge and skills

- Data storage
- Data capture
- Data cleaning
- Data reduction
- Data modelling

Outcome 3

Describe the principal tools used to manipulate big data.

Knowledge and skills

- Hadoop and associated tools (MapReduce/Yarn, Pig, Hive)
- Programming languages (R, Python)
- NoSQL databases

Outcome 4

Use tools to carry out data analysis.

Knowledge and skills

- Predictive analysis
- Data visualisation

Evidence Requirements for this Unit

Learners will need to provide evidence to demonstrate their Knowledge and/or Skills across all Outcomes.

The Evidence Requirements for this Unit will take two forms:

1 Evidence of cognitive competence (for Outcomes 1, 2 and 3).
2 Evidence of practical competence (for Outcome 4).

The evidence of cognitive competence will be the definitions, descriptions and explanations required for Outcomes 1, 2 and 3. The evidence of practical competence will be application of data analysis techniques to a specific problem required for Outcome 4.

Evidence is normally required to for **all** of the knowledge and skills in every Outcome. This means that every knowledge and skills statement must be evidenced. However, sampling may be used in a specific circumstance (see below). For Outcome 4, it is sufficient to apply data analysis techniques to **one** problem.

Evidence must be produced under controlled conditions. However, it is permissible for some evidence to be produced without direct supervision from the assessor. In this case, the evidence must be authenticated. The *Guide to Assessment* provides further advice on methods of authentication. Evidence of authentication must be provided.

There are **no** time limitations on the production of evidence (but see exception below). The evidence may be produced at any time during the life of the Unit.

Sampling is permissible when the evidence for Outcomes 1, 2 and 3 is produced by a test of knowledge and understanding. The test may take any form (including oral) but must be supervised, unseen and timed. The contents of the test must sample broadly and proportionately from the contents of Outcomes 1, 2 and 3 with approximately equal weighting for each Outcome.

The evidence of practical competence (Outcome 4) may relate to a real or fictitious problem. Learners must apply data analysis techniques to a problem.

The Guidelines on Approaches to Assessment (see the Support Notes section of this specification) provides specific examples of instruments of assessment.

Unit Support Notes

Unit title: Data Science (SCQF level 8)

Unit Support Notes are offered as guidance and are not mandatory.

Guidance on the content and context for this Unit

The general context for this Unit is to introduce the principles and practice of data science to learners who already have a basic knowledge of big data. When potentially complex technological or statistical topics are addressed these should be covered in high-level terms.

The Unit should be delivered in an appropriate context for the learners and reflect their vocational and personal interests. For example, learners with an interest and/or background in business should be taught in this context. However, all learners must be exposed to a variety of applications of data science and not just those directly relevant to their vocational interests.

Throughout this Unit it is vital to present the implications of data science in a balanced way, neither over-emphasising the opportunities nor threats posed by this technology.

Outcomes 1, 2 and 3 provide a theoretical under-pinning to the subject of data science. Outcome 4 provides an opportunity to apply this knowledge.

Many useful resources for this Unit can be found online. One useful starting point is the list of Free Data Science Courses at
http://datascienceacademy.com/free-data-science-courses/

An excellent introductory textbook can be downloaded from **http://jsresearch.net/**

A personal, portable Hadoop environment that comes with several interactive Hadoop tutorials can be found at **http://hortonworks.com/products/hortonworks-sandbox/**.

Hadoop for Dummies by Dirk deRoos (John Wiley & Sons, ISBN: 1118607554) gives detailed instructions on how to set up a personal Hadoop environment.

Outcome 1

This knowledge-based Outcome looks at the main areas of operation of data scientists, namely data architecture, data acquisition, data analysis and data archiving.

Data architecture: data scientists can advise data architects on how data should be routed and organised to support subsequent analysis, visualisation and presentation.

Data acquisition: data scientists specify how data should be collected and represented to support analysis and presentation.

Data analysis: data scientists are involved in summarising data, using data samples to make inferences about the population and visualisation of the data by presenting it in graphs, tables or other formats.

Data archiving: data scientists assist in the preservation of collected data in a form that maximises reusability. This can be challenging as it is difficult to anticipate all of the future uses of the data.

Outcome 2

This knowledge-based Outcome examines the techniques used to store and manage big data, focusing on data storage, data capture, data cleaning, data reduction and data modelling.

Data storage: due to the volume of the data and its unstructured or poly-structured nature, big data presents specific data storage challenges. Potential solutions include DAS (Direct Attached Storage), clustered NAS (Network Attached Storage) and object storage.

Data capture: Choosing which data to capture is an important factor, as are the amount of data to be captured and its rate of change.

Data cleaning: Data is often “dirty” due to obsolete, inaccurate or missing information. It must be cleaned up to avoid costly mistakes.

Data reduction: the amount of data needs to be reduced in order to make it comprehensible. The first stage normally involves the use of descriptive statistics. This may be followed by predictive analysis.

Data modelling: This is used to communicate logical and semantic concepts about the data under analysis. It can involve the use of entity-relationship diagrams and similar techniques, but can also include data flows, process models and other types of models.

Outcome 3

This Outcome looks at the major tools used to manipulate big data, including Hadoop and associated tools (MapReduce/Yarn, Pig, Hive), programming languages (R, Python) and NoSQL databases.

All three areas should be covered, but one area may be selected for coverage in greater depth, depending on the tools selected for use in the next Outcome.

Outcome 4

This practical Outcome requires learners to carry out data analysis, using tools selected from those covered in the previous Outcome. Areas covered should include predictive analysis and data visualisation.

This Unit should be delivered using a learner-centred, participative and practical approach.

The Outcomes may be delivered in the order they have been written. They have been written with a learning sequence in mind.

Guidance on approaches to delivery of this Unit

While the exact time allocated to this Unit is at the discretion of the centre, the notional design length is 80 hours. The suggested time distribution is as follows:

Outcome 1: 10 hours
Outcome 2: 10 hours
Outcome 3: 20 hours
Outcome 4: 40 hours

Guidance on approaches to assessment of this Unit

Evidence can be generated using different types of assessment. The following are suggestions only. There may be other methods that would be more suitable to learners.

Centres are reminded that prior verification of centre-devised assessments would help to ensure that the national standard is being met. Where learners experience a range of assessment methods, this helps them to develop different skills that should be transferable to work or further and higher education.

Assessment could be carried using a multiple-choice test of knowledge and understanding for Outcomes 1−3 and a practical assignment for Outcome 4.

The multiple-choice test may take any form (including oral) but must be supervised, unseen and timed. The contents of the test must sample broadly and proportionately from the contents of Outcomes 1, 2 and 3 with approximately 25% of the questions from Outcome 1, 25% from Outcome 2 and 50% from Outcome 3. The test should have a total of 40 questions. A suitable duration could be 90 minutes.

The evidence of practical competence (Outcome 4) may relate to a real or fictitious problem. The practical assignment for this Outcome should require learners to carry out data analysis, using tools selected from those covered earlier. Areas covered should include predictive analysis and data visualisation. This need not be done under supervision, but should be authenticated. This could be done by oral questioning of the learner and/or observation of some aspects of the learner's work.

Opportunities for e-assessment

E-assessment may be appropriate for some assessments in this Unit. By e-assessment we mean assessment which is supported by Information and Communication Technology (ICT), such as e-testing or the use of e-portfolios or social software. Centres which wish to use e-assessment must ensure that the national standard is applied to all learner evidence and that conditions of assessment as specified in the Evidence Requirements are met, regardless of the mode of gathering evidence. The most up-to-date guidance on the use of e-assessment to support SQA's qualifications is available at **www.sqa.org.uk/e-assessment**.

Opportunities for developing Core and other essential skills

This Unit provides opportunities to develop some of the following Core Skills:

- *Information and Communication Technology (ICT)* (SCQF level 6)
- *Numeracy* (SCQF level 6).

Several aspects of the Core Skills in *Information and Communication Technology (ICT)* may be addressed in this Unit. There are opportunities to start software, enter and edit data, locate and extract information, apply a complex search strategy, evaluate information, and present information.

All aspects of the Core Skills in *Numeracy* may be addressed in this Unit. There are opportunities to analyse situations to identify relevant data and relationships, decide which operations to carry out, use numerical and statistical theory, extract, analyse and interpret information, identify significant features in complex graphical information, and select an appropriate graphical form.

History of changes to Unit

Version	Description of change	Date

SQA acknowledges the valuable contribution that Scotland's colleges have made to the development of SQA Advanced Qualifications.

FURTHER INFORMATION: Call SQA's Customer Contact Centre on 44 (0) 141 500 5030 or 0345 279 1000. Alternatively, complete our **Centre Feedback Form**.

General information for learners

Unit title: Data Science (SCQF level 8)

This section will help you decide whether this is the Unit for you by explaining what the Unit is about, what you should know or be able to do before you start, what you will need to do during the Unit and opportunities for further learning and employment.

The purpose of this Unit is to provide an introduction to the theory and practice of data science. The Unit is for those who already have some knowledge of big data and wish a deeper understanding of the techniques used to process it.

The Unit covers a mix of theory and practice. The theoretical content includes the concepts behind data science, the techniques used to store and manage big data and the tools used to process it. The practical content relates carrying out data analysis.

The first Outcome looks at the main areas of operation of data scientists, namely data architecture, data acquisition, data analysis and data archiving.

The second Outcome examines the techniques used to store and manage big data, focusing on data storage, data capture, data cleaning, data reduction and data modelling.

The third Outcome looks at the major tools used to manipulate Big Data, including Hadoop and associated tools (MapReduce/Yarn, Pig, Hive), programming languages (R, Python), noSQL databases.

The final Outcome requires you to carry out data analysis, using tools selected from those covered in the previous Outcome. Areas covered should include predictive analysis and data visualisation.

The Unit seeks to build on existing knowledge of this emerging discipline so that you can appreciate its actual and potential uses in a range of contexts.