Higher National Unit Specification

General information

Unit title: Data Engineering (SCQF level 9)

Unit code: J4YC 36

Superclass: CA

Publication date: September 2020

Source: Scottish Qualifications Authority

Version: 01

Unit purpose

The purpose of this unit is to allow learners gain an understanding on the concepts, application and technologies of data engineering.

This is a **specialist unit** designed for learners wanting to understand and apply the concepts, principles and technologies around data engineering. Learners should be familiar with basic concepts in data and repositories, and ideally have experience using a form of database technologies (either SQL or No SQL). Previous programming experience, in an appropriate language, is assumed.

The unit covers: data engineering principles, data storage architectures and models, database and data-repository technologies (cloud and non-cloud based), data processing frameworks, big data technologies, data warehousing and its application, ETL process, data quality, lineage, monitoring, and implementing schedule data processing jobs.

On the completion of this unit, learners may progress to J4YD 36 *Machine Learning* at SCQF level 9 or J4YA 36 *Statistics for Data* at SCQF level 9.

Outcomes

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

1 Explain the concepts behind data engineering.
2 Describe data storage, architecture and implementation patterns.
3 Explain data processing frameworks for data engineering.
4 Apply data engineering techniques to a problem.

Higher National Unit Specification: General information (cont)

Unit title: Data Engineering (SCQF level 9)

Credit points and level

2 Higher National Unit credits at Scottish Credit and Qualifications Framework (SCQF) level 9: (16 SCQF credit points at SCQF level 9)

Recommended entry to the unit

While entry is at the discretion of the centre, it is recommended that learners possess programming and data analysis skills before commencing this unit. The unit presumes well developed analysis skills, and a familiarity with computer programming in a high level language such as Python or a functional language (such as DAX).

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.

This specialist unit is intended for learners who wish to pursue a career in data science in a data programming capacity. Learners are presumed to have previous experience of data analysis and computer programming before undertaking this unit.

This unit is an ideal progression for learners who have completed units such as J4Y6 35 *Working with Data* at SCQF level 8, J27J 35 *Computer Programming* at SCQF level 8, and J4YD 35 *Programming for Data* at SCQF level 8.

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**.

Higher National Unit Specification: Statement of standards

Unit title: Data Engineering (SCQF level 9)

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

Explain the concepts behind data engineering.

Knowledge and/or skills

- Definition of data engineering and its purpose
- Role of data engineering within data science
- Data engineering major functions
- Data engineering process
- Data engineering tools and techniques
- Data flow
- Data models
- Software engineering principles
- Security concepts
- Trends in data engineering including data ethics

Outcome 2

Describe data storage, architecture and implementation patterns.

Knowledge and/or skills

- Data modelling and storage architecture (structured and unstructured data)
- Traditional databases and big data (data lakes) technologies
- Main data engineering architectures (edge and platform/enterprise)
- Cloud platforms for data architecture and storage (data warehouse)
- Cloud technology toolbox for data engineering functions
- Cloud data implementation patterns (end to end)
- Costing and maintenance of cloud-data environments

Outcome 3

Explain data processing frameworks for data engineering.

Knowledge and/or skills

- Batch processing
- Stream processing
- Interactive processing (online processing)
- Real-time processing
- Parallel processing

Outcome 4

Apply data engineering techniques to a problem.

Knowledge and/or skills

- Architecting distributed systems
- Creating and automating data pipelines
- Architecting data stores
- Validating data quality checks, tracking data lineage, and working with data pipelines
- Governance considerations in data engineering including security and scalability
- System testing
- Solution delivery

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 Knowledge evidence
2 Product evidence

The knowledge evidence relates to outcomes 1–3. Evidence is required for all knowledge and/or skills statements within these outcomes. The amount of evidence may be the minimum required to infer competence. The evidence may be produced over an extended period of time, in lightly controlled conditions.

Higher National Unit Specification: Statement of standards (cont)

Unit title: Data Engineering (SCQF level 9)

The **knowledge evidence** may be sampled when testing is used. In this case, the evidence must be produced under controlled conditions in terms of location, timing and access to reference materials. The sampling frame must cover all outcomes (1–3) but not all knowledge/skills statements; however, the majority of the knowledge/skills must be sampled (at least once) in every instance. The sampling frame must always include the following:

1 Data engineering major functions.
2 Describe types of data modelling and storage architecture.
3 Identify difference between traditional databases and big data (data lakes) technologies and their purposes.
4 Describe data engineering architectures and associated data flows.
5 Explain cloud technologies in data engineering.

The knowledge evidence may be written or oral or a combination of these. Evidence may be captured, stored and presented in a range of media (including audio and video) and formats (analogue and digital). Special consideration should be given to digital formats.

The **product evidence** will relate to outcome 4. It will demonstrate that the learner can apply data engineering techniques to a practical problem. The problem must have sufficient scale and complexity to require an engineering solution. The evidence will demonstrate that the learner has the skills to deliver practical work in the form of a final data engineering project, in a selected platform (cloud or non-cloud), which includes:

- creating and automating data pipelines
- creating data architectures and data stores
- performing data transformations
- creating a clean and high-quality database that could be used for data analysis
- testing the solution
- delivering the solution

The product evidence may be produced over the life of the unit, under loosely controlled conditions (including access to reference materials). When evidence is produced in loosely controlled conditions it must be authenticated. The *Guide to assessment* provides further advice on methods of authentication.

The SCQF level of this unit (level 9) provides additional context on the nature of the required evidence and the associated standards. Appropriate level descriptors should be used when making judgements about the evidence.

The support notes section of this specification provides specific examples of instruments of assessment that will generate the required evidence.

Higher National Unit Support Notes

Unit title: Data Engineering (SCQF level 9)

Unit support notes are offered as guidance and are not mandatory.

While the exact time allocated to this unit is at the discretion of the centre, the notional design length is 80 hours.

Guidance on the content and context for this unit

The first part of this guidance relates to all outcomes. Subsequent parts relate to specific outcomes.

This unit is intended to provide learners with basic concepts to understand in more depth the main definitions and architectures around data engineering, and how to apply them in the design and development of data engineering projects for various purposes. Learners will be expected to have previous knowledge on big data.

Learners will have the opportunity to understand the concepts around data engineering, the key challenges and trends around data engineering and the major functions required in data engineering.

Learners will be provided with guidance and gain understanding on the various data modelling and storage architecture definitions around structured and unstructured data, the differentiation between traditional databases and big data technologies, modern data engineering architecture, such as edge and platform and enterprise, and the use of cloud technologies for the purpose of data engineering.

Additionally, learners will be given guidance on the various data processing frameworks currently available, and how to apply the concepts learnt in this unit, including ETL data processing, data and quality checks for the design and development of a data engineering project.

Please note that the following guidance, relating to specific outcomes, does not seek to explain each knowledge/skills statement, which is left to the professionalism of the teacher. It seeks to clarify the statement of standards where it is potentially ambiguous. It also focuses on non-apparent teaching and learning issues that may be over-looked, or not emphasised, during unit delivery. As such, it is not representative of the relative importance of each knowledge/skill.

At the time of writing, this unit does not lead to recognition by a professional body. It provides some underpinning knowledge for H8W8 34 *Big Data.*

It is to be noted that these outcomes are not intended to be delivered as separate elements of the unit (see *Guidance on approaches to delivery*).

Higher National Unit Support Notes (cont)

Unit title: Data Engineering (SCQF level 9)

Outcome 1: This outcome is intended to provide a broad overview of data engineering and its purpose; the main concepts around data engineering; identifying the role of data engineering within data science, key challenges and trends around data engineering, such as:

- Challenges: Data siloed and difficult to integrate, frontline users unable to generate meaningful data, hidden high costs and operational difficulties with data collection, ineffective end-user training and capabilities in data analytics.
- Trends: Comprehensive end-to-end architecture, more agile IT architecture and operations, move of data collection, processing and storage to the cloud, new trends and application, such as AI and machine learning; data ethics.

It includes the description of major data engineering functions:

- Identifying source data
- Transporting data from sources
- Understanding data and metadata
- Transforming data
- Transporting cured data to new repositories
- Documenting process for repeatability

Outcome 2: This outcome focuses on the description and practical applications of various types of data engineering storage and architecture, and concepts around these themes (cloud and non-cloud based).

The areas that should be covered under this outcome are:

- Data modelling and storage architecture (structured and unstructured data)
- Traditional databases (eg SQL server, Oracle, MySQL, PostgreSQL) and big data (eg Hadoop, Hive, Apache Spark, MongoDB) technologies
- Main data engineering architectures (edge and platform/enterprise)
- Cloud platforms for data architecture and storage (data warehouse) eg AWS, MS Azure
- Cloud technology toolbox for data engineering functions
- Cloud data implementation patterns (end to end)
- Costing and maintenance of cloud-data environments

It is important that as part of this unit outcome, a special focus is given to the design, setup and implementation of Cloud technology data environments (eg AWS, Microsoft Azure), with practical demonstration on how to choose the right data architecture design for a data problem; describing cloud toolbox available for implementing various data engineering processes; costing (moving, storing data) and maintenance of cloud-data environments. Practical demonstrations on setting up cloud data architecture and data engineering patterns to resolve data-related problems are also important.

Higher National Unit Support Notes (cont)

Unit title: Data Engineering (SCQF level 9)

Outcome 3: This outcome covers the understanding of the various data processing frameworks and their implementation (in a cloud or non-cloud platform). For tackling this outcome, learners will be introduced to concepts supporting each of the data processing frameworks, their purpose, performance and architectures, and in which scenarios they are applicable, including their importance in the design of distributed systems. It is not intended that this should be in any way an exhaustive list of such processing frameworks, and the following should be presented as the core ones within data engineering:

- Batch Processing
- Stream Processing
- Interactive Processing (Online Processing)
- Real-Time Processing
- Parallel Processing

Outcome 4: This outcome builds on the knowledge of the data engineering process in automation of data pipelines, including ETL process. This should include activities such as entity extraction and data normalisation algorithms that are used in the data transformation process.

This outcome should cover the understanding and process around data, which includes data quality, tracking data lineage, management of metadata, and important considerations when working with data pipelines. Additional considerations around data engineering concepts, best practices and technologies used in maintenance and governance, will be explained as part of this outcome, including, cybersecurity policies and practices, data engineering cost models and common trade-offs, scalability, and monitoring tools and data-related activities.

The outcome covers the understanding of the steps needed for the implementation of a data engineering project (lifecycle) which are (but not limited to):

- creating and automating data pipelines
- creating data architectures and data stores
- performing data transformations
- creating a clean and high-quality database that could be used for data analysis
- testing the solution
- delivering the solution

Guidance on approaches to delivery of this unit

A suggested distribution of time, across the outcomes, is:

Outcome 1: 15 hours
Outcome 2: 25 hours
Outcome 3: 15 hours
Outcome 4: 25 hours

It is anticipated that the required concepts will be introduced by the teacher and reinforced with appropriate examples, especially in the description of concepts, case studies of actual potential use or real-case scenarios are advised.

Summative assessment may be carried out at any time. However, when testing is used (see evidence requirements) it is recommended that this is carried out towards the end of the unit (but with sufficient time for remediation and re-assessment). When continuous assessment is used (such as the use of a web log), this could commence early in the life of the unit and be carried out throughout the duration of the unit.

There are opportunities to carry out formative assessment at various stages in the unit. For example, formative assessment could be carried out on the completion of each outcome to ensure that learners have grasped the knowledge contained within it. This would provide assessors with an opportunity to diagnose misconceptions and intervene to remedy them before progressing to the next outcome.

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 out using:

- A selected response test that covers the knowledge and understanding for outcomes 1, 2 and 3
- An assignment that covers the knowledge and understanding for outcomes 1, 2 and 3
- A set of practical tasks that covers the practical competence for outcome 4

Each selected response question could be structured as four options (one key) with a pass mark of 60% for the whole test. Use should be made of scenario type questions to assess the learner's competency in distinguishing various data storage architecture, technologies and the data processing frameworks. The test could consist of a relatively high number of questions (30 or 40, for example), lasting an hour, which would span outcomes 1, 2 and 3, and sample all of the knowledge statements (including at least one question for each statement).

The assignment could require the learner to research and present evidence that they can describe the various types of data engineering modelling and architectures. The evidence must cover knowledge of data engineering processing frameworks as learned in this unit. The evidence should be in the learner's own words, and all references to the sources of information included in the evidence. This assignment would be undertaken over a defined period of time.

Higher National Unit Support Notes (cont)

Unit title: Data Engineering (SCQF level 9)

The practical tasks could be carried out over an extended period of time. They would allow the learner to demonstrate competence in implementing a data engineering project that should include creating and automating data pipelines, creating data architectures and data stores, performing data transformations, creating a clean and high-quality database that could be used for data analysis, testing the solution and delivering the solution. The set of practical tasks must cover all of the practical competences set out in outcome 4.

A more contemporary approach to assessment would involve the use of a web log (blog) to record learning (and the associated activities) throughout the life of the unit. The blog could provide knowledge evidence (in the descriptions and explanations). The blog should be assessed using defined criteria to permit a correct judgement about the quality of the digital evidence. In this approach to assessment, every knowledge and skill must be evidenced; sampling would not be appropriate.

Formative assessment could be used to assess learners' knowledge at various stages throughout the life of the unit. An ideal time to gauge their knowledge would be at the end of each outcome. This assessment could be delivered through an item bank of selected response questions, providing diagnostic feedback to learners (when appropriate).

If a blog is used for summative assessment, it would also facilitate formative assessment since learning (including misconceptions) would be apparent from the blog, and intervention could take place to correct misunderstandings on an on-going basis.

It is important to ensure that work submitted by a learner is their own. The risk of malpractice is greater when you do not have the opportunity to observe learners carrying out assessment activities. There are various web-based services that can detect plagiarism, but the following strategies can also be effective in authenticating learners' work:

- questioning
- write-ups under controlled conditions
- witness testimony
- use of personal logs
- personal statements produced by your learners

The use of case studies which require learners to include information from their own experience can also help to reduce plagiarism. You should ensure that learners are clear about how to access resources, especially from the internet, how to reference the material they use, and the extent to which they may confer with others or seek support.

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**.

Higher National Unit Support Notes (cont)

Unit title: Data Engineering (SCQF level 9)

Opportunities for developing Core and other essential skills

The unit provides opportunities to develop some of the following Core Skills:

- Communication at SCQF level 6
- Information and Communication Technology (ICT) at SCQF level 6
- Problem Solving at SCQF level 6
- Numeracy at SCQF level 6

Learners are expected to make use of software to design and code solutions to data-engineering learning scenarios. They will address several components of the Core Skill in *Information and Communication Technology (ICT)* in so doing.

Learners will be required to make decisions about which data-related architecture, technologies, tools and frameworks are to be applied to given problems, and to implement the appropriate solution. They will address several components of the Core Skill in *Problem Solving* in so doing.

Learners will present evidence from their assignment in ways that demonstrate their understanding and communicate key concepts. They will address several components of the Core Skill in *Communication* in so doing.

Learners will handle data, perform transformations, understand data relationships and decide which operations to carry out on it. They will also encounter a range of numerical and statistical concepts and address several components of the Core Skill in *Numeracy* in so doing.

History of changes to unit

Version	Description of change	Date

Unit template: June 2017

General information for learners

Unit title: Data Engineering (SCQF level 9)

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 you with an understanding on the concepts, application and technologies of data engineering. It is intended for anyone who has an appreciation of the importance of data to their personal and professional life and wishes to understand better the fundamental concepts on which its application is based. It would be beneficial if you had already undertaken J4Y6 35 *Working with Data* (SCQF level 8).

The unit comprises four outcomes and develops your knowledge and understanding of the concepts behind data engineering along with some practical competence in using software tools that perform activities around data, from the design to the implementation of data engineering projects.

Some of the topics covered in this unit include:

- Concept of data engineering, and its purpose
- Key challenges around data and its processing, and technological trends to solve some of the data engineering challenges
- Major functions in data engineering, and the activities required to fulfil them, including identifying data sources, transporting data, understanding of data and metadata, transforming data and documenting the process for maintenance and repeatability
- Application of a range of data engineering and architecture concepts and technologies (cloud and non-cloud base), with a special focus on cloud data technology as an important platform used for various data engineering functions and projects
- Various types of data modelling and storage architecture (structured and unstructured)
- Difference between traditional databases (eg SQL server, Oracle, MySQL) technologies and big data technologies (eg Hadoop, Hive, Apache Spark, MongoDB)
- Definition of main data engineering architectures (edge and platform/enterprise), to understand and implement cloud data platforms (eg MS Azure, AWS) and their respective data engineering functions, cloud data warehouses and their architecture
- Data processing frameworks and their implementation, eg Batch Processing, Stream Processing, Interactive Processing (Online Processing), etc
- Automation of data pipelines including ETL (extract-transform-load) process
- Data extraction and data normalisation
- Data maintenance and governance, data quality, tracking data lineage, management of metadata, and important considerations when working with data pipelines

Assessment will likely be through a range of assessment methods, most of which will be highly practical in nature, including assignments and case studies.

This unit will also provide opportunities for you to enhance your competence in Core Skills such as *Information and Communication Technology (ICT), Numeracy* and *Problem Solving*.