Next Generation Higher National Unit Specification

Data Engineering (SCQF level 9)

Unit code:	J6CL 49
SCQF level:	9 (16 SCQF credit points)
Valid from:	session 2023–24

Prototype unit specification for use in pilot delivery only (version 1.0) September 2023

This unit specification provides detailed information about the unit to ensure consistent and transparent assessment year on year.

This unit specification is for teachers and lecturers and contains all the mandatory information required to deliver and assess the unit.

This edition: September 2023 (version 1.0)

Unit purpose

This specialist unit is designed for learners who want 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 of using a database technology (either SQL or NoSQL). Previous programming experience, in an appropriate language, is assumed.

While entry is at your centre's discretion, we recommend that learners have programming and data analysis skills before starting this unit. The unit presumes learners have 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).

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
- extract, transform and load (ETL) process
- data quality
- lineage
- monitoring
- implementing schedule data processing jobs

On the completion of this unit, learners can progress to further units in data science at SCQF level 9 and higher.

Unit outcomes

Learners who complete this unit can:

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

Evidence requirements

Learners must provide both knowledge and product evidence.

Knowledge evidence

Evidence should relate to outcomes 1, 2 and 3, and is required for all knowledge and skills statements within these outcomes. The amount of evidence can be the minimum required to infer competence and produced over an extended period, under supervised conditions.

You can sample evidence when using a question paper. In this case, learners must provide the evidence under controlled conditions in terms of location (supervised), timing (limited) and access to reference materials (not allowed). Sampling must cover outcomes 1 to 3 but not all knowledge and skills statements; however, you should sample most of the knowledge and skills at least once in every instance.

Sampling must always include the following:

- describing data engineering major functions
- describing types of data modelling and storage architecture
- identifying the difference between traditional databases and big data (data lakes) technologies and their purposes
- describing data engineering architectures and associated data flows
- explaining cloud technologies in data engineering

Evidence can be written or oral or a combination of these. It can be captured, stored and presented in a range of media (including audio and video) and formats (analogue and digital). You should give special consideration to digital formats.

Product evidence

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

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

Learners can produce evidence over the duration of the unit, in unsupervised conditions (including access to reference materials). When evidence is produced in unsupervised conditions it must be authenticated. The ’Approaches to assessment’ section provides further advice on methods of authentication and specific examples of assessment.

The standard of evidence should be consistent with the SCQF level of this unit.

You should use appropriate level descriptors when making judgements about the evidence.

Knowledge and skills

The following table shows the knowledge and skills covered by the unit outcomes:

Knowledge	Skills
Learners should understand: ♦ the definition of data engineering and its purpose ♦ the role of data engineering within data science ♦ data engineering major functions ♦ the data engineering process ♦ data engineering tools and techniques ♦ data flow ♦ data models ♦ software engineering principles ♦ security concepts ♦ trends in data engineering including data ethics ♦ data modelling and storage architecture (structured and unstructured data) ♦ the traditional databases and big data (data lakes) technologies ♦ the main data engineering architectures (edge and platform/enterprise) ♦ the cloud platforms for data architecture and storage (data warehouse) ♦ the cloud technology toolbox for data engineering functions ♦ the cloud data implementation patterns (end-to-end) ♦ the costing and maintenance of cloud data environments ♦ batch processing ♦ stream processing ♦ interactive processing (online processing) ♦ real-time processing ♦ parallel processing ♦ architecting distributed systems ♦ governance considerations in data engineering including security and scalability	Learners can: ♦ create and automate data pipelines ♦ architect data stores ♦ validate data quality checks, track data lineage, and work with data pipelines ♦ create a data engineering solution ♦ test the solution ♦ deliver the solution

Meta-skills

Throughout the unit, learners develop meta-skills to enhance their employability in the data science sector.

Self-management

This meta-skill includes:

- focusing: sorting, attention, filtering
- integrity: ethics
- adapting: adaptability, self-learning, resilience
- initiative: independent thinking, decision making

Social intelligence

This meta-skill includes:

- communicating: receiving information, listening, giving information
- feeling: social conscience
- collaborating: team working and collaboration

Innovation

This meta-skill includes:

- curiosity: information sourcing, problem recognition
- creativity: imagination, idea generation, visualising, maker mentality
- sense-making: pattern recognition, holistic thinking, synthesis, opportunity recognition, analysis
- critical thinking: deconstruction, logical thinking, judgement, computational thinking

Literacies

Throughout this unit, learners have opportunities to develop their literacy skills.

Numeracy

Numeracy is developed in some of the knowledge and skills, including:

- data manipulation
- data transformation
- data analysis

Communication

Communication is developed in some of the knowledge and skills, including:

- ethical aspects
- using open source ETL tools
- communication and collaboration skills

Digital

The knowledge and skills of this unit significantly contribute to learners’ digital skills.

Delivery of unit

This unit provides specialist knowledge and skills in data engineering, and you can deliver it with other units in the group award.

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

We suggest the following distribution of time:

Outcome 1 — Explain the concepts behind data engineering
(15 hours)
Outcome 2 — Describe data storage, architecture and implementation patterns
(25 hours)
Outcome 3 — Explain data processing frameworks for data engineering
(15 hours)
Outcome 4 — Apply data engineering techniques to a problem
(25 hours)

These outcomes are not intended to be delivered as separate elements of the unit.

You should introduce the required concepts of the unit and provide appropriate examples, especially in the description of concepts. We advise using case studies of actual potential use or real-case scenarios.

If you deliver the unit as part of a group award, we recommend that you teach and assess within the subject area of the group award.

If you assess evidence for outcomes on a sample basis, you must teach the whole of the content listed in the 'Knowledge and skills' section, so it is available for assessment. Learners should not know in advance the items you are assessing, and you should sample different items on each assessment occasion.

Additional guidance

The guidance in this section is not mandatory.

Content and context for this unit

This unit provides learners with an understanding of 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 should have previous knowledge on big data.

You should give learners the opportunity to understand the:

- concepts around data engineering
- key challenges and trends around data engineering
- major functions required in data engineering

Provide guidance so they can 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, platform and enterprise
- the use of cloud technologies for the purpose of data engineering

Additionally, you should guide learners on the various data processing frameworks currently available, and how to apply the concepts they have learned in this unit. This includes ETL data processing, and data and quality checks for the design and development of a data engineering project.

The following guidance relating to specific outcomes, does not explain each knowledge and skills statement — this is at your discretion. It clarifies the statement of standards where it is potentially ambiguous. It also focuses on non-apparent learning and teaching issues that may be over-looked, or not emphasised, when delivering the unit. As such, it is not representative of the relative importance of each knowledge and skill.

At the time of writing, this unit does not lead to recognition by a professional body, however, it provides some underpinning knowledge for the Big Data unit at SCQF level 7.

Explain the concepts behind data engineering (outcome 1)

This provides a broad overview of data engineering and its purpose; the main concepts around data engineering; identifying the role of data engineering within data science; and 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 in the cloud
  - new trends and application, such as artificial intelligence (AI) and machine learning
  - data ethics

This outcome provides a 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

Describe data storage, architecture and implementation patterns (outcome 2)

This 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 covered include:

- data modelling and storage architecture (structured and unstructured data)
- traditional databases (for example, SQL server, Oracle, MySQL, PostgreSQL) and big data (for example, Hadoop, Hive, Apache Spark, MongoDB) technologies
- main data engineering architectures (edge and platform or enterprise)
- cloud platforms for data architecture and storage (data warehouse), for example, Amazon Web Services (AWS), Microsoft (MS) Azure
- cloud technology toolbox for data engineering functions
- cloud data implementation patterns (end-to-end)
- costing and maintenance of cloud data environments

In this outcome, it is important that you focus on:

- the design, setup and implementation of cloud technology data environments (such as AWS, MS Azure), with practical demonstration on how to choose the right data architecture design for a data problem
- describing the cloud toolbox available for implementing various data engineering processes
- costing (moving, storing data)
- the 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.

Explain data processing frameworks for data engineering (outcome 3)

This covers understanding the various data processing frameworks and their implementation (in a cloud or non-cloud platform). In this outcome, you introduce learners 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. This is not an exhaustive list, but you should present the following as the core processing frameworks in data engineering:

- batch processing
- stream processing
- interactive processing (online processing)
- real-time processing
- parallel processing

Apply data engineering techniques to a problem (outcome 4)

This builds on the knowledge of the data engineering process in automation of data pipelines, including the ETL process. You 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
- important considerations when working with data pipelines

Additional considerations around data engineering concepts, best practices and technologies used in maintenance and governance are explained as part of this outcome, including:

- cybersecurity policies and practices
- data engineering cost models and common trade-offs
- scalability
- monitoring tools and data-related activities

This outcome also covers understanding the steps needed to implement 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

Approaches to assessment

Evidence can be generated using different types of assessment. The following are suggestions, however, there may be other methods that could be more suitable for your learners.

We suggest the following approaches to assessment:

- Multiple-choice questions that cover the knowledge for outcomes 1, 2 and 3.
  - Each multiple-choice question could be structured as four options (one key), with a pass mark of 60 per cent for the exam. You should use scenario question types when assessing learners identifying various data storage architecture, technologies and the data processing frameworks. The exam could last 60 minutes and could include approximately 30 or 40 questions, covering outcomes 1, 2 and 3. This exam would sample all the knowledge statements, including at least one question for each statement.
- An assignment that covers the knowledge for outcomes 1, 2 and 3.
  - Learners could research and present evidence, showing that they can describe the different types of data engineering modelling and architectures. Evidence must cover the knowledge of data engineering processing frameworks they have learned in this unit, in their own words, with all references to information sources included in the evidence. This assignment would be created over an extended period.
- A set of practical tasks that cover the practical competencies for outcome 4. The practical tasks could be carried out over an extended period. Learners could 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
  - delivering the solution

A more contemporary approach to assessment would be learners writing a blog. It would record their learning and the associated activities for this unit, and could provide knowledge evidence in the descriptions and explanations. You should assess the blog using defined criteria to ensure that you record a correct judgement on the quality of the digital evidence. In this approach to assessment, you must have evidence for every knowledge and skill and sampling is not appropriate.

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

You can use formative assessment to assess learners' knowledge at different points of this unit. An ideal time to check learners' knowledge would be at the end of each outcome. You can deliver this assessment through an item bank of multiple-choice questions, providing feedback to learners (when appropriate).

When using continuous assessment (such as a blog), this can commence early in the life of the unit and continue throughout the duration of the unit.

You can carry out summative assessment at any time. However, when using a question paper, we recommend that you carry it out towards the end of the unit, but with sufficient time for remediation and re-assessment (see 'Evidence requirements' section).

Using a blog for summative assessment also helps formative assessment, because learning (including misconceptions) is clear, and you can intervene to correct misunderstandings on an ongoing basis.

It is important to check and ensure that learners' submitted work is their own. The risk of malpractice is greater when you do not 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

Using case studies that require learners to include information from their own experience can also help to reduce plagiarism. You should ensure that learners are clear about:

- accessing resources (especially from the internet)
- referencing the material they use
- the extent to which discussing or seeking support from others is appropriate

Equality and inclusion

This unit is designed to be as fair and as accessible as possible with no unnecessary barriers to learning or assessment.

You should take into account the needs of individual learners when planning learning experiences, selecting assessment methods or considering alternative evidence.

Guidance on assessment arrangements for disabled learners and/or those with additional support needs is available on the assessment arrangements web page.

Information for learners

Data Engineering (SCQF level 9)

This information explains:

- what the unit is about
- what you should know or be able to do before you start
- what you need to do during the unit
- opportunities for further learning and employment

Unit information

This unit follows on from the Data Engineering unit at SCQF level 8 and is for learners who wish to build on their knowledge and understanding of data engineering.

It is for anyone who has an appreciation of the importance of data to their personal and professional life and wishes to have a greater understanding of the fundamental concepts on which its application is based. It is beneficial if you have already completed the Working with Data unit at SCQF level 8.

The unit provides you with an understanding of the concepts, application and technologies of data engineering. It also 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 include:

- the concept of data engineering, and its purpose
- the key challenges around data and its processing, and technological trends to solve some of the data engineering challenges
- the 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
- the application of a range of data engineering and architecture concepts and technologies (cloud and non-cloud based), with a special focus on cloud data technology as an important platform used for various data engineering functions and projects
- the various types of data modelling and storage architecture (structured and unstructured)
- the difference between traditional databases (for example, SQL server, Oracle, MySQL) technologies and big data technologies (for example, Hadoop, Hive, Apache Spark, MongoDB)
- the definition of main data engineering architectures (edge and platform or enterprise), to understand and implement cloud data platforms (for example, MS Azure, AWS) and their respective data engineering functions, cloud data warehouses and their architecture
- data processing frameworks and their implementation, for example batch processing, stream processing, interactive processing (online processing)

- automation of data pipelines including ETL 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

The unit covers a wide range of meta-skills. The meta-skills you develop cover self-management, social intelligence, and innovation. For example, you improve your self-management skills by making decisions based on data and you cover ethical aspects of data engineering, contributing to your communication skills. You also develop your numerical and data skills throughout this unit.

On completion, can progress to further units in data science at SCQF level 9 and higher.

Administrative information

Published: September 2023 (version 1.0)

Superclass: RB

History of changes

Version	Description of change	Date

Note: please check SQA's website to ensure you are using the most up-to-date version of this document.