Next Generation Higher National Unit Specification

Data Engineering (SCQF level 8)

Unit code:	J6CD 48
SCQF level:	8 (24 SCQF credit points)
Valid from:	session 2023–24

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

Unit purpose

This specialist unit introduces learners to the principles and practice of data engineering. It is for learners with an interest in pursuing a career in data science and is particularly suitable for those studying a Higher National Diploma (HND) in Data Science and those who are interested in progressing to university or who want to pursue a career in data engineering or data science. It is designed to deliver the specialist knowledge and skills that are required to be a data engineer.

No previous experience of data engineering is required but learners should be familiar with data analysis and data programming. They can evidence this by completing the Working with Data unit at SCQF level 8 and the Programming with Data unit at SCQF level 8.

The unit builds on knowledge of data analysis and programming. It introduces the engineering principles and practices involved in building end-to-end data processing systems.

Learners are introduced to:

- the principles of data engineering and software engineering
- data engineering processes and tools, including cloud technologies
- appreciation of problems that data engineering can address

On completing this unit, learners:

- understand the role of a data engineer
- understand data architecture, data storage and data processing, particularly as it applies to big data
- have further knowledge of database technologies such as SQL and NoSQL
- understand data warehousing and cloud migration
- appreciate the implications of machine learning for data engineering

Learners can progress to the Data Engineering unit at SCQF level 9.

Unit outcomes

Learners who complete this unit can:

1 explain the principles of data engineering
2 describe the extract, transform and load (ETL) process
3 explain data architectures and data models
4 explain the mechanisms for moving and processing of data
5 apply data engineering principles and practices to a problem

Evidence requirements

Learners must provide both knowledge and product evidence.

Knowledge evidence

Learners must demonstrate the knowledge evidence for all outcomes.

Evidence may be sampled for outcomes 1, 2, 3 and 4 when testing is used. When used, testing must be carried out under supervised conditions and in a controlled environment. Learners cannot access reference materials under exam conditions. Sampling, on all occasions, must include outcomes 1 to 4, but not every knowledge statement associated with each outcome needs to be sampled.

Product evidence

Evidence should relate to outcome 5. It must demonstrate that learners can apply the principles of data engineering to at least one problem. Their evidence must demonstrate that they can:

- construct a data pipeline
- build a data warehouse
- use tools to extract, transform and load data
- create data models
- write code to analyse data
- create a dashboard to present data

The problem that learners choose to apply the principles of data engineering to must acquire data from several sources (in a variety of formats) and require significant transformation and modelling before analysis. The dataset must be large and complex and need an engineered solution. Their analysis does not need to be sophisticated, but they must demonstrate robust programming methods. Learners must present the results from their analyses in a clear, engaging, easy-to-use, interactive format.

Evidence can be produced over an extended period, under lightly controlled conditions and without assistance. Authentication is required.

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 ♦ the functions of data engineering and its place in data science ♦ the responsibilities of a data engineer ♦ data ethics ♦ data bias and algorithmic bias ♦ programming languages for data engineering ♦ programming methods including structured programming ♦ software engineering principles ♦ algorithms and data structures for data engineering ♦ SQL and NoSQL ♦ machine learning ops ♦ data mining ♦ the data engineering process, and its tools and techniques ♦ ETL process ♦ data sources and data formats ♦ data management and data security ♦ data schemas and data models ♦ data pipelines ♦ distributed databases ♦ data lakes and data warehousing ♦ parallel computing ♦ cloud computing for data storage and data processing ♦ processing systems including stream and real-time ♦ online analytical processing (OLAP) and online transactional processing (OLTP)	Learners can: ♦ visualise and create data pipelines ♦ construct a data warehouse ♦ develop programming skills about writing large, complex systems with high volumes of data ♦ use ETL tools to extract data from various sources ♦ use ETL tools to clean large volumes of data ♦ use ETL tools to transform, map, and manipulate data in various formats ♦ create data models from schemas ♦ write database queries ♦ write computer programs to perform ETL tasks ♦ analyse and visualise data ♦ create automated solutions to analyse data

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 extraction
- data manipulation
- data transformation
- data analysis
- data visualisations

Communication

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

- ethical aspects of data engineering
- using open source ETL tools
- communication and collaboration
- presentation skills

Digital

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

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 120 hours.

We suggest the following distribution of time:

Outcome 1 — Explain the principles of data engineering
(20 hours)
Outcome 2 — Describe the extract, transform and load process
(20 hours)
Outcome 3 — Explain data architectures and data models
(20 hours)
Outcome 4 — Explain the mechanisms for moving and processing data
(20 hours)
Outcome 5 — Apply data engineering principles and practices to a problem
(40 hours)

You should complete learning and teaching in outcome order.

Learners should be familiar with data analysis and data programming. They can show this by completing the Working with Data unit at SCQF level 8 and the Programming with Data unit at SCQF level 8.

Learners with previous knowledge can advance quickly through this unit. For example, learners with previous experience of databases and programming may already have some of the required knowledge and skills and may not need additional teaching. They may have some of the required evidence (see 'Evidence requirements' section). However, this needs to be authenticated.

Learners require access to a range of digital technologies, such as:

- personal computers
- tablets and/or smartphones
- internet
- a range of data engineering tools

Additional guidance

The guidance in this section is not mandatory.

You should help learners to understand and implement the concepts around data engineering and the key tasks that a data engineer performs by understanding the requirements of a data consumer.

You should provide a broad overview of topics that is consistent with the SCQF level of the unit. You should not assume previous knowledge of data engineering, although learners should be familiar with data analysis and computer programming.

You should teach learners the basic concepts of:

- data engineering
- the key challenges and trends around data engineering
- various data engineering processes, tools and techniques
- the major tasks that a data engineer performs

Help learners to gain an understanding of the various types of data modelling and data storage systems, and the differentiation between traditional databases and big data technologies. Explain the difference between an operational database and a data warehouse, and the use of cloud technologies for data engineering.

Additionally, you should give learners guidance on the various data processing frameworks that are currently available. Teach them how to apply the concepts learnt in this unit, including ETL data processing, and data and quality checks for the design and development of a data engineering project.

It is important that learners experience using a variety of data engineering tools during the delivery of the course. For example, you can explore more than one ETL tool to cover the skills requirement, and most ETL tools have similar architecture.

The following guidance relating to specific outcomes, does not explain each knowledge and skills statement — this is at your discretion.

Explain the principles of data engineering (outcome 1)

This provides a basic understanding of the important principles used in data engineering. It gives a broad outline of the definition and functions of data engineering including, but not limited to:

- how data engineering is an aspect of data science
- the common responsibilities of a data engineer, such as communicating effectively across organisational, technical and political boundaries, and responding to challenges
- translating data into valuable insights that inform decisions, and developing data services to meet user needs
- learning about programming languages for data engineering that may include at least one programming language such as R, Python, Java, used by data engineers

- learning about various algorithms and data structures for data engineering
- machine learning ops
- a brief overview of data mining
- a brief overview of different types of data models
- explaining how to develop, test and maintain an organisation’s data infrastructure including data warehouses and data pipelines
- using programming languages and tools to complete a data engineering project

Challenges in data engineering include, but are not limited to:

- metadata management
- integrating multiple data systems
- ensuring quality of data outputs in terms of completeness, consistency, conformity, accuracy, integrity and timeliness of data

They also include operational difficulties in a data engineering process.

Describe the extract, transform and load (ETL) process (outcome 2)

This covers the in-depth knowledge of the extract, transform and load (ETL) process used by data engineers. The ETL process is an important operational aspect of a data warehouse.

Sources from which data are extracted during the ETL process can be OLTP or legacy systems and can be in various formats, such as spreadsheets, text documents and web pages.

This outcome also includes data ethics (you should include the UK Data Ethics Framework when using data in public sector), data bias and algorithmic bias.

Explain data architectures and data models (outcome 3)

This provides an in-depth understanding of various common data architectures and data models used in creating various types of data engineering storage. This includes, but is not limited to:

- data lake architecture
- single-tier, two-tier and three-tier data warehouse architecture

Various other components of a data warehouse architecture you should include are:

- operational systems
- transactional databases
- the concept of meta-data
- data marts
- query tools

Data lake architecture including:

- ingestion tier
- insights tier
- the Hadoop Distributed File System (HDFS)
- distillation tier
- processing tier
- unified operations tier

You should include data lake concepts such as data ingestion, data storage, data governance, and security.

This outcome provides further knowledge of database technologies such as SQL, NoSQL and distributed databases, cloud migration, and the implications of machine learning for data engineering.

Explain the mechanisms for moving and processing of data (outcome 4)

The focus should be on the commonly used data processing mechanisms in a data engineering system. Other basic types of data processing mechanisms and concepts should include, but are not limited to:

- transaction processing
- distributed processing
- real-time processing
- multiprocessing
- workflow
- monitoring
- batch processing
- stream processing
- OLAP
- OLTP

You should present real-world data processing pipeline examples in a problem-solving context.

Apply data engineering principles and practices to a problem (outcome 5)

This builds on the knowledge and skills of the data engineering process learnt in the first four outcomes of this unit. It covers understanding the steps needed for the implementation of at least one of the following:

- a data warehouse using the ETL process
- a data lake using the ETL process
- an ETL pipeline

There are many open source ETL tools available to help you perform practical activities to capture the skills requirements, such as Talend Studio, Apache Kafka to store and publish a stream of records, Tableau prep, Stich, SQL Server Integration Services (SSIS).

The treatment of data visualisation can be light, but learners must gain an understanding of tools that they can use to visualise the data after completion of the ETL process. An automated solution to analyse data should be generated.

When delivering this unit, it is important that you take every opportunity to introduce data engineering concepts in a real-world context. This should be in the form of case studies, where data engineering solutions can solve real-world problems.

Approaches to assessment

Centres can create any assessments that satisfy the evidence requirements for this unit.

A traditional approach to assessment could involve an exam for knowledge evidence (outcomes 1 to 4) and a practical assignment or project for product evidence (outcome 5). The exam could use a number of extended response questions sampling outcomes 1 to 4, ensuring that every outcome is included. For example, the exam could contain ten extended response questions, worth 10 marks each, out of 100 marks with a pass mark of 50 marks. The assignment could require learners to design and implement a data pipeline, including warehousing, and use that system to analyse their data.

An alternative approach could involve learners creating and maintaining a blog to record their learning activities during this unit, generating sufficient evidence to satisfy the requirements for outcomes 1 to 4. Product evidence could be generated using an assignment as described above (outcome 5).

Another approach could require learners to maintain a portfolio of their work throughout the unit, to evidence their cognitive and practical competencies.

Authentication is required when the evidence is produced under lightly controlled conditions.

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 8)

This section 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 specialist unit introduces you to the principles and practice of data engineering. No previous knowledge of the subject is needed, but you should have previous experience of data analysis and data programming before attempting this unit.

During this unit you learn how to use a wide range of ETL tools and write database queries. You also develop your data programming skills with an emphasis on formal programming methods and software engineering principles.

On completing this unit, you should be able to:

1 explain the principles of data engineering
2 describe the extract, transform and load (ETL) process
3 explain data architectures and data models
4 explain the mechanisms for moving and processing of data
5 apply data engineering principles and practices to a problem

You are assessed using a combination of methods. For example, your knowledge can be assessed in an exam and your practical skills can be assessed through an assignment or project. Alternatively, you may be asked to create and maintain a blog or portfolio.

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.

As this unit provides specialist knowledge and skills in data engineering, you can go on to study more advanced units, such as the Data Engineering unit at SCQF level 9.

Administrative information

Published: June 2023 (version 1.0)

Superclass: CB

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.