Next Generation Higher National Unit Specification

Big Data (SCQF level 8)

Unit code: J6CB 48

SCQF level: 8 (16 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 unit introduces learners to the theory, principles and practice of big data. It is a specialist unit for learners who require a deeper understanding of this emerging discipline, and its actual and potential uses in a range of contexts.

Before starting this unit, learners should have previous experience of computing and data analysis, such as the Digital Skills unit at SCQF level 7 and the Working with Data unit at SCQF level 7 and/or level 8. Although no previous experience of big data is required, learners would benefit from completing the Big Data unit at SCQF level 7.

This unit covers a range of topics relating to big data, including:

- the reasons for the growth of data
- contemporary applications of big data
- the key technologies that enable big data processing and analysis
- ethical and social implications

A significant proportion of time is devoted to handling and processing big data from a public source, including some basic descriptive analytics.

On completion of this unit, learners have an appreciation of the importance of big data in contemporary societies, its applications in a range of contexts, and the technologies that enable big data to be collected, stored and analysed. They gain skills in acquiring, processing and analysing big data. Learners can progress to more specialised units at SCQF level 8 or level 9.

Unit outcomes

Learners who complete this unit can:

1 describe the characteristics that distinguish big data
2 explain the challenges that big data brings for data processing methods
3 explain the architecture and technological requirements of big data
4 acquire, load and analyse big data to describe its key characteristics
5 investigate a contemporary application of big data
6 critique the gathering and use of big data in contemporary business and society

Evidence requirements

Learners must provide knowledge and product evidence.

The knowledge evidence takes the form of an investigation into a contemporary application of big data. Learners must carry out this investigation without assistance, and their report must include the:

- reasons for the big data application
- value of the application
- data sources and data volumes
- types of data
- types of analytics deployed
- data architecture and data pipelines
- big data technologies employed
- limitations of the application
- ethical implications of the application

The investigation must relate to a contemporary, real-life application of big data technologies and techniques. It must be a substantial technical report into the application, incorporating narratives, processes, schematics and other visualisations, where appropriate.

Leaners should generate the product evidence using one or more software tools to acquire data from an external source, perform validation and integrity checks, create transformed fields, and store the data in a datastore (warehouse) located in the cloud. This dataset should have more than 100 000 records and more than 10 fields.

Learners should extract a dataset matching a given criteria from this warehoused data. This dataset should have more than 10 000 records and more than 10 fields. They should use a data analysis tool to produce a given set of descriptive statistics, and they should display specific information extracted from this dataset in a dashboard.

The evidence consists of a portfolio containing:

- the logical data model for the ETL (extract, transform and load) process
- the set of rules for data cleansing and transformation
- screenshots of the various intermediate processes in producing the dataset to be stored
- a brief report of the conclusions from the descriptive statistics generated by analysis
- screenshots of the dashboard created from the dataset and a description of its contents

The evidence can be produced over an extended period of time in lightly controlled conditions, or holistically generated in conjunction with other units within a group award. Authentication is required when the evidence is produced in lightly controlled conditions.

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: ♦ key drivers for the growth of data ♦ terminology and characteristics of big data ♦ the three Vs of big data (volume, velocity, variety) ♦ the big data value chain and lifecycle ♦ data warehousing and the ETL process ♦ types of data (unstructured, semi-structured, structured, metadata) ♦ data formats (CSV, XML, JSON, Parquet) ♦ RDBMS and NoSQL databases and their limitations for big data ♦ big data technologies, including Hadoop ♦ public sources of large datasets ♦ data cleansing, filtering and conversion ♦ types of data analytics, including descriptive, predictive and prescriptive ♦ contemporary applications of big data, including user-behaviour analytics and predictive analytics ♦ ethical and social implications of big data, including privacy and current legislation regarding big data collection, storage and use ♦ the implications for individuals and society in the use of big data	Learners can: ♦ create a logical data map for an ETL process ♦ import data from a source and process it for storage ♦ perform checks for data integrity and validity ♦ perform operations to clean and transform data according to data rules ♦ extract a dataset from a data store for analytical purposes ♦ produce descriptive statistics for a dataset ♦ use data analytics software to produce a dashboard of information from a dataset ♦ produce a technical report on a big data project

Meta-skills

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

Self-management

This unit prompts learners to understand the basics of data and how it is loaded into a professional tool. This requires them to identify, organise, filter and sort data in a logical manner, and then select relevant analytical tools to effectively understand the information generated. Learners should also understand how to query any provided statistics to ensure that they are not skewed, thereby giving a misleading impression. The big data project also needs to understand the limits presented by timescales.

Social intelligence

This unit prompts learners to identify relevant big data from a variety of sources and integrate these effectively. They should use this analysis to determine how to represent the information to an audience in an appropriate format. They should identify various exploration and visualisation tools to help them conclude what this information means and how it can be used to make effective decisions.

You should encourage learners to examine the data supplied for completeness and accuracy, and to take into account any null values entered due to lack of available information.

Innovation

Learners are equipped with the necessary skills to generate a variety of outputs to assess what type would be most appropriate for the problem domain. They would do this from data gathered using a number of different sources, including different types of visualisation.

This unit allows learners to understand the output as an effective means of representing the underlying data, and suggest what trends can be predicted from the information generated. This may involve a brief introduction to artificial intelligence (AI) and predictive analysis using datasets, for example, a simple binary classification with a supplied script and sample dataset.

Literacies

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

Numeracy

Numeracy is developed throughout the unit through activities that include data gathering, data transforming, data analysis and data exploration, and data visualisation.

Communication

Learners develop their communication skills throughout the unit, and particularly when researching a big data application and preparing an extensive report explaining and critiquing the application. They also develop their ability to explain technical details.

Digital

This unit contributes to digital skills. Learners broaden their digital skills as they develop their understanding of the progression of data through the big data lifecycle. This includes appropriate collection methods, through to analysis, and eventually in using it to make decisions. They use digital tools to acquire and process data, to analyse data, and present their findings through visualisations.

Delivery of unit

This unit provides foundational knowledge and skills and is likely to be delivered before other units in a group award. Nonetheless, there may be opportunities to deliver it concurrently with more specialised units.

While the exact time allocated to this unit is at your centre's discretion, the notional design length is 80 hours. One possible approach is to distribute the available time as follows:

Outcome 1 — Describe the characteristics that distinguish big data
(10 hours)
Outcome 2 — Explain the challenges that big data brings for data processing methods
(10 hours)
Outcome 3 — Explain the architecture and technological requirements of big data
(10 hours)
Outcome 4 — Acquire, load and analyse big data to describe its key characteristics
(25 hours)
Outcome 5 — Investigate a contemporary application of big data
(15 hours)
Outcome 6 — Critique the gathering and use of big data in contemporary business and society
(10 hours)

Learners with pre-existing digital skills may accelerate through the unit. For example, learners with previous experience of data analysis or big data may already have some of the required knowledge and skills, and may not require additional teaching and learning. These learners may have some (or all) of the required evidence (see 'Evidence requirements' section).

Additional guidance

The guidance in this section is not mandatory.

Your level of treatment should be sufficient to allow learners to understand the relevant stages in big data processes, and the role that data architecture and technology play in managing big data.

Learners should experience a range of tools to perform the ETL process, including a distributed data store. They should also become familiar with data formats used in large public databases, such as CSV, XML and Parquet, and know the limitations of traditional database management when dealing with high data volumes (both SQL and NoSQL).

You could use spreadsheet software to demonstrate the principles of data ingestion, cleansing and transformation. This requires learners to become familiar with spreadsheet functionality beyond a basic level, to include selecting, copying, deleting, filtering, sorting and other typical operations. You should also expose learners to some datasets that cannot be processed in this way.

You should underline the value of data for decision making and innovation, introduce learners to the data analysis process, and apply it to a dataset using a contemporary data analysis tool. Options for this practical work include using a programming language and libraries, such as Python, NumPy and Plotly; tools from the Apache suite of big data applications, such as Spark and Storm; or tools such as Google Dataflow, Google Data Studio and Power BI.

You should also introduce basic analytics in this part of the unit, including common statistical measures, such as location and spread, and correlation.

You could deliver outcomes 1, 2 and 3 by drawing on case studies that exemplify the concepts that characterise big data, and tracing over time the growth of data and the technology that enables its processing. This gives learners a foundation on which to conduct an in-depth investigation of a big data application for outcome 5. They can build up evidence over the course of the unit, preferably through an e-portfolio.

The practical work for outcome 4 is intended to expose learners to a large dataset (at least 100 000 records), which requires a big data technology to ingest, process and store. You should provide guidance on an appropriate toolset, along with appropriate instruction in its application. Learners with programming skills might prefer to adopt a programming approach using custom libraries for big data applications, or use a tool like those from Amazon Web Services (AWS), Azure or Google. A good approach, that gives learners an understanding of industry practice, is to introduce them to tools from the Apache suite, such as Hadoop, Spark and Storm.

As part of outcome 4, learners could carry out data analysis using a tool that is more amenable to learner installation and implementation, such as Tableau Public, Google Data Studio or Power BI. As before, learners with programming skills could use Python or RStudio along with specialist libraries, such as NumPy, SciKit and SciPy, and Plotly, to achieve the necessary analysis and visualisation.

By expanding on the ethical implications of the chosen application for their outcome 5 case study, learners can incorporate the evidence for outcome 6 into their report.

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

Big Data (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 unit introduces you to the theory and practice of big data. No previous knowledge of big data or statistics is required, but you should have good numeracy skills and be familiar with the data analysis process.

Big data techniques can be applied in the analysis of human behaviour and natural phenomena, and are becoming widely used in various fields, ranging from business to health and government. This unit seeks to explain what big data is, how it might affect you, and how you might use it in your job. It demonstrates how organisations currently use big data techniques to mine personal data, using the knowledge gained to either generate revenue or drive policy and/or business decisions.

This unit is designed to build your confidence in using digital technology for a wide range of personal and vocational purposes, and to introduce you to the theory, practice and applications of big data in modern society and business. You gain an understanding of the value chain for big data and the structure of the big data pipeline.

You use a range of digital tools to acquire a large dataset, process it and store it, and then use data analysis to generate conclusions about the dataset. You also learn some terminology used in programming and data analysis. You can include the use of spreadsheets and databases, or other tools required to analyse a dataset, and then present your findings in the form of tables or charts. The presentation of conclusions derived from data is a key skill required by most employers nowadays, irrespective of what area or field of specialisation they operate in.

You are assessed by means of an investigative report that you prepare on your choice of a specific application of big data. This should include a critique of the application, and demonstrate your competence in downloading a large dataset and applying tools to check the data, adjust errors and generate new data. You analyse this dataset and produce a brief summary showing the outcomes of your analysis, including a dashboard comprising a few representations of your data.

This unit covers a wide range of meta-skills, like self-management, social intelligence and innovation, along with other literacies. For example, you develop self-management skills as you make decisions based on data and how that data should best be represented. You build on your numerical, communication and digital literacies throughout this unit — particularly your digital literacy.

Administrative information

Published: June 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.