IfATE
Shaping skills training

Data engineer

Key information

Reference: ST1386
Version: 1.0
Level: 5
Typical duration to gateway: 24 months
Typical EPA period: 4 months
Maximum funding: £19000
Route: Digital
Date updated: 11/12/2023
Approved for delivery: 11 December 2023
Lars code: 746

EQA provider: Ofqual
Example progression routes:
Digital and technology solutions specialist (integrated degree),
Digital and technology solutions specialist (integrated degree)
Review: this apprenticeship will be reviewed in accordance with our change request policy.

Contents

Details of the occupational standard

Occupation summary

This occupation is found in a wide range of public and private sector organisations who work with large data sets including Government departments, NHS, financial and professional services, IT companies, retail and sales and education providers.

The purpose of the occupation is to build systems that collect, manage, and convert data into usable information for data scientists, data analysts and business intelligence analysts to interpret. A data engineer's main aim is to make data accessible and valid so that an organisation can use it to evaluate and optimise their performance. The role of the data engineer is pivotal to any organisation; it ensures that data pipelines are established to support data scientists and other business stakeholders.

A data engineer will build and implement data flows to connect operational systems, and re-engineer manual data flows to enable scalable and repeatable use. They integrate, support and manage the build of data streaming systems, writing extract transform and load scripts that perform in line with business requirements.

They are responsible for providing high quality, transparent data that enables effective governance and smart business decisions. They will analyse the performance indicators of the data systems that provide clean, regular, and accurate data. A data engineer will understand how data and an organisation's data architecture is essential to business outcomes.

A data engineer will be able to gather requirements for data solutions, and they demonstrate and articulate data solutions to stakeholders in a way that can be easily understood. Data engineering encompasses a range of activities from collecting data to employing various data processing frameworks, including but not limited to ETL (Extract, Transform, Load), and collaborating with data scientists and other data-centric roles. Data engineers may work in an office or work remotely depending on the sector they work in and location of the employer.

In their daily work, an employee in this occupation will work autonomously or collaboratively with clients, in the business and or data team. A data engineer will work with data analysts, Data scientists and data architects and liaise with other teams and internal and external stakeholders to ensure their data requirements are captured and managed to the specified standard. They will also work closely with machine learning engineer (Ops), software engineers, software developers and technology teams.

An employee in this occupation will be responsible for completing their own work to specification, , ensuring they meet set deadlines. A data engineer contributes towards, engineering designs, plans, execution and evaluation working to time, cost and quality targets. They deliver to the product roadmap and are responsible for meeting quality requirements and working in accordance with health and safety and environmental considerations. They will work according to organisational procedures and policies, to maintain security and compliance.

Typical job titles include:

Data engineer

Occupation duties

DUTY	KSBS
Duty 1 Build and optimise automated data systems and pipelines considering data quality, description, cataloguing, data cleaning, validation, technical documentation and requirements.	K1 K2 K3 K4 K5 K6 K7 K8 K9 K10 K11 K12 K26 S1 S2 S3 S4 S5 S6 S7 S8 S13 B1 B2 B3
Duty 2 Integrate, support and manage data using standalone, distributed and cloud-based platforms. To ensure efficient, sustainable and effective provision of data storage solutions.	K3 K4 K10 K11 K13 K14 K15 S1 S2 S3 S7 S9 S10 B2 B4
Duty 3 Support the identification and evaluation of opportunities for data acquisition and data enrichment.	K10 K11 K16 K17 S11 S12 B2 B3
Duty 4 Select and use appropriate tools to process data in any format, such as structured and unstructured data and in any mode of delivery, such as streaming or batching. Adapt to legacy systems as required.	K18 K19 K20 S9 S14 S15 S16 B4
Duty 5 Ensure resilience is built into data products against business continuity and disaster recovery plans, and document change management to limit service outages. Support and respond to incidents through the application of technology and service management best practice including configuration, change and incident management.	K21 K22 K23 S10 S18 S19 S20 S21 B2 B3 B5
Duty 6 Analyse requirements, research scope and options and present recommendations for solutions to stakeholders.	K12 K30 S2 S3 S11 S22 S23 S27 B2 B3 B4
Duty 7 Support the implementation of prototype or proof-of-concept data products within a production environment	K6 K8 K24 K30 S17 S22 S24 B3 B4
Duty 8 Maintain data solutions as continually evolving products, to service the organisation, user or client requirements. Collaborate with technical support teams and stakeholders from	K6 K25 K30 S1 S2 S3 S11 S12 S21 S22 S25

implementation to management.	B1 B3 B4
Duty 9 Working within compliance and contribute towards data governance, organisational policies, standards, and guidelines for data engineering.	K3 K4 K6 K10 K11 S5 S13 S17 B3
Duty 10 Monitor the data system to meet service requirements to enable solutions such as data analysis, dashboards, data products, pipelines, and storage solutions.	K27 S18 S19 S26 B3 B5
Duty 11 Keep up to date with engineering developments to advance own skills and knowledge.	K28 K29 S12 S28 S29 B6

KSBs

Knowledge

K1: Processes to monitor and optimise the performance of the availability, management and performance of data product.

K2: Methodologies for moving data from one system to another for storage and further handling.

K3: Data normalisation principles and the advantages they achieve in databases for data protection, redundancy, and inconsistent dependency.

K4: Frameworks for data quality, covering dimensions such as accuracy, completeness, consistency, timeliness, and accessibility.

K5: The inherent risks of data such as incomplete data, ethical data sources and how to ensure data quality.

K6: Software development principles for data products, including debugging, version control, and testing.

K7: Principles of sustainable data products and organisational responsibilities for environmental social governance.

K8: Deployment approaches for new data pipelines and automated processes.

K9: How to build a data product that complies with regulatory requirements.

K10: Concepts of data governance, including regulatory requirements, data privacy, security, and quality control. Legislation and its application to the safe use of data.

K11: Data and information security standards, ethical practices, policies and procedures relevant to data management activities such as data lineage and metadata management.

K12: How to cost and build a system whilst ensuring that organisational strategies for sustainable, net zero technologies are considered.

K13: The implications of financial, strategic and compliance regarding to security, scalability, compliance and cost of local, remote or distributed solutions.

K14: The uses of on-demand Cloud computing platform(s) in a public or private environment such as Amazon AWS, Google Cloud, Hadoop, IBM Cloud, Salesforce and Microsoft Azure.

K15: Data warehousing principles, including techniques such as star schemas, data lakes, and data marts.

K16: Principles of data, including open and public data, administrative data, and research data including the value of external data sources that can be used to enrich internal data. Examples of how business use direct data acquisition to support or augment business operations.

K17: Approaches to data integration and how combining disparate data sources delivers value to an organisation.

K18: How to use streaming, batching and on-demand services to move data from one location to another.

K19: Differences between structured, semi-structured, and unstructured data.

K20: Types and uses of data engineering tools and applications in own organisation.

K21: Policies and strategies to ensure business continuity for operations, particularly in relation to data provision.

K22: Technology and service management best practice including configuration, change and incident management.

K23: How to undertake analysis and root cause investigation.

K24: Processes for evaluating prototypes and taking them to implementation within a production environment.

K25: The lifecycle of implementing data solutions in a business, from scoping, though prototyping, development, production, and continuous improvement.

K26: Data development frameworks and approved organisational architectures.

K27: The principles of descriptive, predictive and prescriptive analytics.

K28: Continuous improvement including how to: capture good practice and lessons learned.

K29: Strategies for keeping up to date with new ways of working and technological developments in data science, data engineering and AI.

K30: The methods and techniques used to communicate messages to meet the needs of the audience.

Skills

S1: Collate, evaluate and refine user requirements to design the data product.

S2: Collate, evaluate and refine business requirements including cost, resourcing, and accessibility to design the data product.

S3: Design a data product to serve multiple needs and with scalability, efficiency, and security in mind.

S4: Automate data pipelines such as batch, real-time, on demand and other processes using programming languages and data integration platforms with graphical user interfaces.

S5: Produce and maintain technical documentation explaining the data product, that meets organisational, technical and non-technical user requirements, retaining critical information.

S6: Systematically clean, validate, and describe data at all stages of extract, transform, load (ETL).

S7: Work with different types of data stores, such as SQL, NoSQL, and distributed file system.

S8: Identify and troubleshoot issues with data processing pipelines.

S9: Query and manipulate data using tools and programming such as SQL and Python. Manage database access, and implement automated validation checks.

S10: Communicate downtime and issues with database access to stakeholders to mitigate the operational impact of unforeseen issues.

S11: Evaluate opportunities to extract value from existing data products through further development, considering costs, environmental impact and potential operational benefits.

S12: Maintain a working knowledge of data use cases within organisations.

S13: Use data systems securely to meet requirements and in line with organisational procedures and legislation.

S14: Identify new tools and technologies and recommend potential opportunities for use in own department or organisation.

S15: Optimise data ingestion processes by making use of appropriate data ingestion frameworks such as batch, streaming and on-demand.

S16: Develop algorithms and processes to extract structured data from unstructured sources.

S17: Apply and advocate for software development best practice when working with other data professionals throughout the business. Contribute to standards and ways of working that support software development principles.

S18: Develop simple forecasts and monitoring tools to anticipate or respond immediately to outages and incidents.

S19: Identify and escalate risks with suggested mitigation/resolutions as appropriate.

S20: Investigate and respond to incidents, identifying the root cause and resolution with

internal and external stakeholders.

S21: Identify and remediate technical debt, assess for updates and obsolescence as part of continuous improvement.

S22: Develop, maintain collaborative relationships using adaptive business methodology with stakeholders such as, business users, data scientists, data analysts and business intelligence teams.

S23: Present, communicate, and disseminate messages about the data product, tailoring the message and medium to the needs of the audience.

S24: Evaluate the strengths and weaknesses of prototype data products and how these integrate within an organisation's overarching data infrastructure.

S25: Assess and identify gaps in existing tools and technologies in respect of implementing changes required.

S26: Identify data quality metrics and track them to ensure the quality, accuracy and reliability of the data product.

S27: Selects and apply sustainable solutions to contribute to net zero and environmental strategies across the various stages of product and service delivery.

S28: Horizon scanning to identify new technologies that offer increased performance of data products.

S29: Implement personal strategies to keep up to date with new technology and ways of working.

Behaviours

B1: Acts proactively and takes accountability adapting positively to changing work priorities, ensuring deadlines are met.

B2: Works collaboratively with stakeholders and colleagues, developing strong working relationships to achieve common goals. Support an inclusive culture and treat technical and non- technical colleagues and stakeholders with respect.

B3: Quality focus that promotes continuous improvement utilising peer review techniques, innovation and creativity to the data system development process to improve processes and address business challenges.

B4: Takes personal responsibility towards net zero and prioritises environmental sustainability outcomes in how they carry out the duties of their role.

B5: Use initiative and innovation to problem solve and trouble shoot, providing creative solutions.

B6: Keeps abreast of developments in emerging, contemporary and advanced technologies to optimise sustainable data products and services.

Qualifications

English and Maths

Apprentices without level 2 English and maths will need to achieve this level prior to taking the End-Point Assessment. For those with an education, health and care plan or a legacy statement, the apprenticeship's English and maths minimum requirement is Entry Level 3. A British Sign Language (BSL) qualification is an alternative to the English qualification for those whose primary language is BSL.

Version log

Version	Change detail	Earliest start date	Latest start date
1.0	Approved for delivery	11/12/2023	Not set