



Level 5
Data Engineer

Higher Apprenticeship

Programme Handbook 2024

Contents

Introduction to the Programme	4
Key Facts	5
Entry Routes to the Programme	6
Programme Delivery	7
Tips for Success	8
Achieving a Distinction	8
The Learning Journal	9
Inductions	9
Module Coverage	10
Module List	10
Module Explainers	11
Data Fundamentals	11
Databases and Data Lakes	13
Programming and Scripting Essentials	15
Networks Essentials and Cyber Security	16
Cloud Engineering	18
Data Collection and Ingestion pt. 1	21
Data Collection and Ingestion pt. 2	23
Data Pipelines	24
Data Responsibilities	28
Production Environment	30
Data Products	34
Data Futures	37
End Point Assessment (EPA) Gateway	39
End Point Assessment (EPA)	40
Assessment Part 1: Project	41
Assessment Part 2 – Professional Discussion	42
Overview of the professional discussion	42
Formative submissions	43
Formative submissions of your work on the programme	43
The role of the portfolio of evidence	44
The STAR technique	45
Real world evidence	45
Evidence - good practice and advice	45
Getting to EPA - recap	46
Unfunded learners	46
KSBs by Assessment Methods (Mapping Table)	47
Project KSBs by theme	53
Professional discussion KSBs by theme	55

Grading basis	58
Grades available	61
Overall EPA grading	61
Re-sits and re-takes	61
Programme roles	62
New Funding Rules 22/23.....	62
Active Learning Rule	62
What constitutes evidence?	62
Off the job training rule	63
What is OTJT	63
List of all Knowledge Skills & Behaviours (KSBs)	64
Knowledge	64
Skills	65
Behaviours	66
Maths and English on the Programme	67
Safeguarding, Prevent and British Values	68
Core Reading List	69
Short Industry Certifications.....	70

Introduction to the Programme

This novel Higher Apprenticeship is delivered via the new Level 5 (L5) Data Engineer apprenticeship standard. The standard, first approved by IFATE in December 2023, has been designed by employers and takes account of the Data Engineering knowledge, skills and behaviours needed in the modern workplace.

We work in collaboration with both you (the **Learner / Apprentice**) and your line manager from the outset to ensure that you have access to relevant opportunities in the workplace. Throughout the programme, you will develop a portfolio which presents evidence of real-life work projects that you have engaged in; personal reflection will play a significant part in this process demonstrating how you have applied knowledge, skills and behaviours associated with the apprenticeship standard. You will be exposed to a wealth of practical work preparing you for your endpoint assessment.

The standard Level 5 Data Engineer Higher Apprenticeship spans 15 months of teaching and learning + a 4-month End Point Assessment (EPA) period, totalling 19 months of learner engagement on the standard programme. If your programme follows an accelerated path, you will receive your timings separately.

The primary objectives of this programme are to:

- Equip the learner with comprehensive knowledge and proficiency in designing, building, evaluating and maintaining sustainable systems that manage data as it is collected, moved, stored, transformed, organised and accessed. Through mastering the knowledge, skills and behaviours related to data products, databases, data lakes, data pipelines, programming, scripting, data governance, metrics monitoring, networks, cyber security, cloud infrastructure, and production environments. As well as communicating and collaborating with technical and non-technical stakeholders.
- Foster a deep understanding of how data engineering competencies and best practices enable a Data Engineer to carry out their professional duties. By equipping the learner with competencies related to data standards, data collection, good data handling practices, data project management, data services provision, continuous integration and deployment, future-proofing data, visualisation, maintaining data quality and responding to issues. Ensuring data service reliability that will allow all other business users to trust the data and derive actionable insights from it.
- Balance theory and practical application, ensuring learners are given ample opportunities for on-the-job learning to complement their off-the-job training. Additionally, completing a portfolio of real-life applications of the knowledge, skills and behaviours learnt on the programme, working ethically, and identifying relevant areas for continuous professional development.

This L5 programme is positioned to prepare the learner to be able to embark on a range of career journeys that require Data Engineering knowledge, skills and behaviours., Such as Data Engineer, Cloud Solutions Architect, Data Architect, Senior Data Engineer, Data Engineering Manager, Chief Technology Officer, Chief Data Officer, Software Engineer in Test, Data / Platform Operations, ETL Developer, Developer Advocate, Data Infrastructure Engineer, Data Services Manager, Big Data Solutions Specialist, DevOps, Data Ops, Database Administrator, Data/Cloud Migration and Integration Specialist, IT/Infrastructure Tester, Forward Deployed Engineer among other similar roles. The programme may also act for some learners as a springboard to L6/L7 Apprenticeships, BSc/BEng Top-Up programmes and integrated MSc/MEng degrees in related areas (requirements apply).

Learners will need to plan for 6-8 hours each week to complete off-the-job (OTJ) guided study and portfolio work. This would normally take place every week on a fixed day of the week.

Key Facts

This is a Level 5 qualification spanning 15 taught months, followed by the endpoint assessment period (EPA) of 4 months, for a total of 19 months of learner engagement, with 487 planned hours of learner engagement, containing:

- 487 hours of off-the-job, guided study, including:
 - 189 live taught hours (webinars delivering both theory and hands-on practicals)
 - 174 hours of interactive online content (videos, programming notebooks, etc.)
 - 4 hours of induction with the tutor at the beginning of the programme, including learning
 - 120 notional hours of off-the-job portfolio work demonstrating how the learner has applied taught concepts in practice
 - The learner builds a portfolio throughout the course, incorporating concepts they learnt
 - This acts as evidence of the knowledge, skills and behaviours required for the EPA
 - Notional hours get credited in chunks during regular progress updates with the learner's coach
 - These are broken down into 5 portfolio "tasks", each requiring 24 hours of notional work

The above are the hours that will be tracked on the programme. There are multiple options for highly motivated students, including those students who are aiming for a Distinction to stretch and extend their learning experience, and this will be explained to learners.

Entry Routes to the Programme

This Level 5 programme is designed to support a diverse range of learners seeking up-to-date Data Engineering skills and knowledge:

- Entry-level learners: individuals new to the areas of Data and Engineering, with computer literacy and a keen interest in exploring the field of Data Engineering, especially:
 - Professionals beginning their professional journeys who have a keen interest in Data Engineering and wish to develop specialised skills in this field from the outset of their careers.
- Continuing learners: individuals proficient in data fundamentals for data analytics or data business operations or in similar fields, but seeking to broaden their knowledge base with Data Engineering best practices and/or re-position themselves within a Data Engineering pathway for their career and/or education goals, especially:
 - Professionals who wish to re-train through practical, work-based learning, such as employees currently working in roles that involve data handling who aim to advance their career in the engineering function.

For any candidate, the following are the **recommended minimum entry requirements**, ensuring that the candidate comes equipped with a generalist understanding of computer technologies underpinning the technical infrastructures of organisations, whilst being familiar with input-output systems for data and their basic use cases:

- (L3) Data Foundations / Data Technician or a related Advanced Apprenticeship qualification, **or**
- An A-Level in Computer Science / Maths or a Sciences / Technology subject, **or**
- An equivalent IT / ICT / Computing / Software / Technology or Engineering qualification with a TQT of 360 hours minimum, **or**
- Completed work experience of 720 hours minimum in an IT / Computing / Software / Technology / Infrastructure or Engineering oriented role, this needs to be attested by the learner's manager. Having two A-Levels or equivalent or five GCSEs grades A–C or equivalent is recommended in this scenario.

This programme is **also suitable** for learners who have attained a qualification at a level that is already higher than the minimum requirement, such as:

- L4 Data Analyst / L4 Business Analyst, or
- Motivated learners with other suitable L4 qualifications wanting to convert into Data Engineering
OR
- Learners with suitable L3 and L4 qualifications wishing to be considered for a Top-Up BSc year in Engineering / Computing / Applied Data, or a related field, upon completion of this programme (learners are advised to ensure that they have enough credits at lower levels prior to starting this programme; learners are advised that apprenticeship programmes do not specify credit points and any conversion of credit requirements will be at the discretion of the institution they are progressing into; learners are informed that progression is not guaranteed; criteria apply)
OR
- Learners with suitable L3 and L4 qualifications wishing to be eligible for a Level 6 (BSc) Digital Technology degree, or pursue a Level 7 (MSc) Digital and Technology Solutions Specialist integrated degree, upon completion of this programme (learners are informed that progression is not guaranteed; criteria apply).

Candidates should be employed in a related field.

Apprentices without Level 2 **English and Maths** will need to achieve this in the first 12 months of the programme. 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.

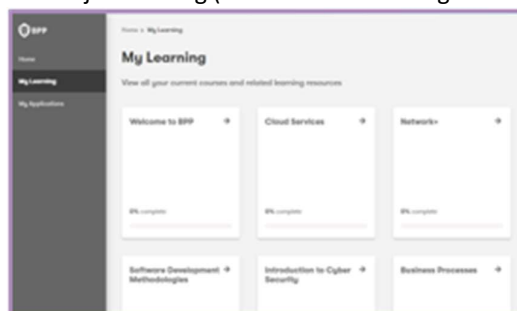
Programme Delivery

BPP hold a rolling cohort start model, where new cohorts commence the programme at regular intervals throughout the Year, usually every three months. Below are the key facts about how the programme is delivered during the 15-month practical period:

- The sessions (also known as webinars, or workshops) are delivered weekly (with a two-week Christmas break).
- The sessions are delivered live using Adobe Connect. The standard duration is 3 hours, with breaks in the middle.
- There will be three “hackathons” spread throughout the programme. On those days, learners will have two sessions – one in the morning, and one in the afternoon. These full-day teaching days provide a unique immersive learning experience. The dates are announced in advance so learners can block them in their calendars.
- Apprentices have a dedicated day of the week for their learning. The majority of the learning regularly takes place on the same day of the week.
- The sessions are designed to be interactive with discussions, group activities, and hands-on exercises guided by the instructor.
- The rest of the content is available online on The Hub website, where each learner has an account.
- Guided technical exercises are provided for each module. Some of them may be completed individually without the instructor present.
- A so-called “spiral curriculum” is embedded throughout the programme, so learners are constantly invited to revisit, reinforce, and build upon learning. There will be an overarching case study that will start out simple, and we will be filling it in with more detail as we progress through the programme.
- The learners are reminded about updating their portfolio, keeping a reflective journal, and documenting their reading and on-the-job training.
- The learners are also reminded to engage with their coach for booking individual coaching sessions.
- There will also be a Skills Scan element, where the learners will be periodically contacted to measure their familiarity with different concepts and skills. This is a learning support practice and is not part of the overall assessment. You will be contacted about Skills Scan separately.

The delivery model provides the learner with interactive online materials available each week to set the scene and prepare the learner for the upcoming teaching sessions. The delivery model has one 3-hour live taught session per week (usually mornings), where the afternoon is guided to be spent practicing with interactive online materials that we provide. The student will use additional time on the day to work on their portfolio.

Learners are required to attend live webinars unless they have a valid reason for non-attendance. Learners are advised to be free for 6-8 hours on their study days to make the most out of their programme, and their managers should be on board with recognising the time commitment required for success. This would normally take place every week on a fixed day of the week. Learners wanting to achieve a distinction are also advised to allocate a smaller chunk of hours spread across their remaining working week to log their on-the-job training (see section Achieving a Distinction).



All of the learner’s programme information, learning resources and key communications will be on The Hub. When you log in, you will see a home page for the programme, and you will be able to access each module. Each module will be clearly structured so that you can see, each week, which topic you are studying and what activities you should be doing before and after your weekly webinar. We monitor usage of The Hub and expect that you will be accessing it regularly during term time. All learners are required to attend live webinars unless they have a reason for absence. All online webinars are recorded and are available for knowledge consolidation, revision and supported learning.

Tips for Success

We encourage a range of learners from technical and non-technical backgrounds to apply to this programme, as long as they find Data Engineering interesting and are willing to commit to exploring this field.

The following advice will be useful to learners who are at the beginning of their journey with Data Engineering:

1. **Set Clear Goals:** Break your learning into small, achievable objectives to track progress and stay motivated.
2. **Embrace Growth:** Believe in your capacity to learn and improve. Every challenge is a chance to grow.
3. **Stay Curious:** Let your curiosity drive you deeper into data engineering topics, making learning an adventure.
4. **Participate in Online Forums:** Join discussions on Kaggle and StackOverflow to learn from others and share your insights. They are great platforms for practical learning and problem-solving.
5. **Learn from Mistakes:** View mistakes as learning opportunities. Reflecting on errors helps you improve.
6. **Use Interactive Tools:** Engage with hands-on tutorials and coding exercises to strengthen your understanding of complex concepts.
7. **Visual Aids:** Simplify tough topics with diagrams and videos, which can help clarify and retain information.
8. **Apply Your Knowledge:** Work on small projects to apply what you've learned, reinforcing your skills.
9. **Manage Your Time:** Allocate dedicated periods for focused learning to enhance efficiency. Stick to regular times for best results.
10. **Seek Feedback:** Regularly ask for feedback from your manager, coach, and peers. It's crucial for identifying strengths and areas to work on.
11. **Document Your Journey:** Keep a record of your learning experiences. It can deepen your understanding and serve as a valuable resource.
12. **Balance Your Efforts:** Ensure a healthy balance between work, learning, and personal well-being to maintain energy and motivation.
13. **Use The Hub:** Engage with The Hub, particularly the forums. They offer a supportive space for questions, discussions, and insights from both peers and experts.
14. **Communicate Regularly:** Keep in touch with your manager and coach about your progress and challenges. Open communication supports your development.
15. **Embrace All Learning Opportunities:** Maximize growth by equally focusing on both on-the-job and off-the-job training, leveraging every chance to learn and apply new skills.

Achieving a Distinction

To achieve a distinction, a motivated candidate would normally incorporate the following habits into their on-the-job learning:

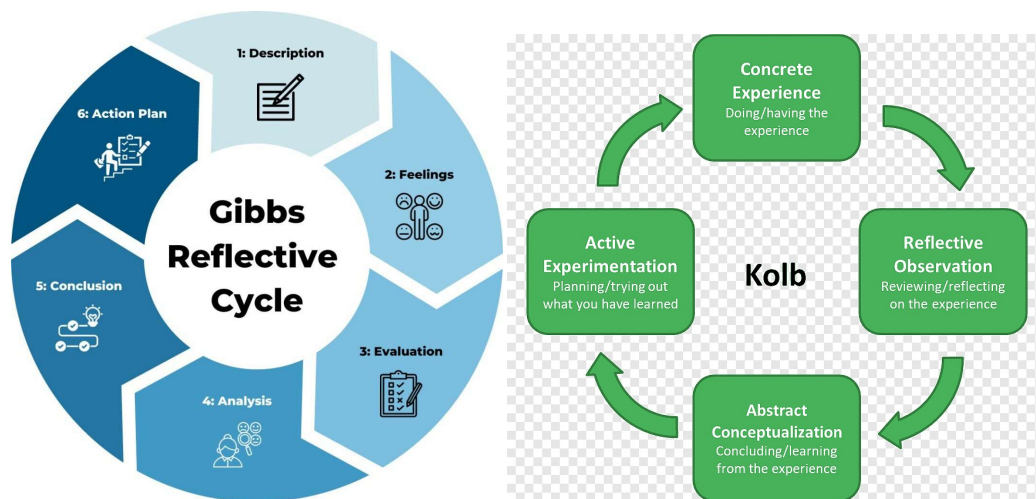
- Reading items from the core reading list (see end of handbook)
- Participating in online communities such as Kaggle and StackOverflow
- Attempting short-course industry certifications from the approved list (see end of handbook)
- Identifying and shadowing relevant in-house staff at your place of work
- Identifying and participating in your in-house short training courses that are relevant to your employer and offered at your place of work

- Collaborating on a variety of data-related projects at work, seek out new opportunities to collaborate
- Participate in the Hub activities marked as stretch and extend activities

For the items listed above, the learner interested in achieving a distinction may want to allocate a smaller chunk of hours spread across their remaining working week to make sure they create a habit of doing these things regularly (no more than 6-7 extra hours). They may also want to log any of the hours spent on the above activities in the optional tracker included as an attachment at the end of this handbook, to ensure that their coach and their manager have a visibility of all their activities. (Please note that this is separate from the OTJ Tracker which is compulsory for all learners and only counts the required core hours).

The Learning Journal

The Level 5 Data Engineer programme contains a mandatory learning journal element that will be explained by your tutor at the beginning of your first module. You are expected to create a living document that you will be updating **throughout your whole Apprenticeship programme** with summaries of what you have learned as well as your reflections on the content and how it applies to your employment.



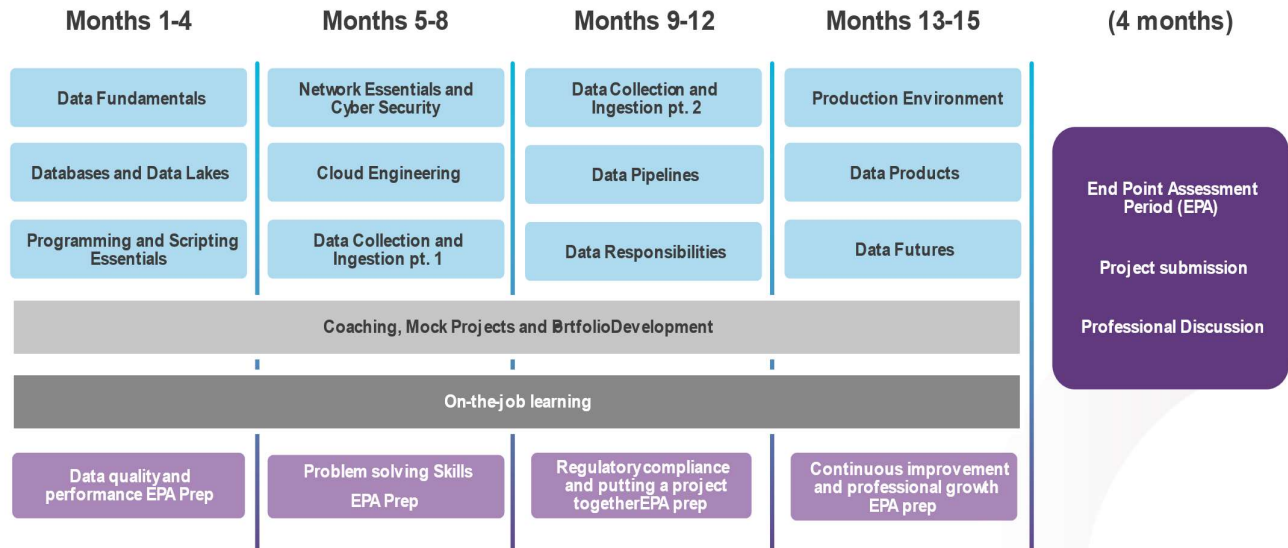
The two approaches recommended for writing your learning journal are the Gibbs Reflective Cycle and Kolb's reflective model of experiential learning. As you can see from the diagrams above, learning is an iterative process, as it requires cycles/repetition of experiencing new skills, knowledge, and behaviours, then evaluating them (taking into account thoughts, feelings and analysis), and using the reflective process to then create action plans and experiments to further your learning, opening you up to new learning experiences.

Please note that the learning journal is separate from the portfolio and the formatives, which are explained later in this Handbook in the EPA section. The learning journal will not attract a formal mark or grade but you may be asked to show your learning journal to either BPP or your employer at any time for the purposes of feedback and support.

Inductions

Inductions are led by the Award Lead. They provide learners and line managers with a programme overview and outline of expectations around what the programme requires of them. Inductions are online in a mix of Teams and Adobe Connect.

Module Coverage



Module List

- Data Fundamentals
- Databases and Data Lakes
- Programming and Scripting Essentials
- Network Essentials and Cyber Security
- Cloud Engineering
- Data Collection and Ingestion pt. 1
- Data Collection and Ingestion pt. 2
- Data Pipelines
- Data Responsibilities
- Production Environment
- Data Products
- Data Futures

Module Explainers

Data Fundamentals

Module	Phase	Weekly Topics	Description
Data Fundamentals	Month 1	<ul style="list-style-type: none">• Fundamentals of the data-driven enterprise: estimating and monitoring the business value of Big Data and small data; data ecosystems in modern organisations: data fabric and data mesh.• Introduction to data quality: data standards, data policies, data principles, structured and unstructured data types and formats, data lineage, identifying relevant metrics.• Introduction to managing data projects and products: collaborating with stakeholders in Agile and Lean data teams; optimising data discovery through data catalogues, data dictionaries and data stewardship principles; overview of data product types and the data lifecycle. Sustainability and net-zero considerations;• Introduction to reliable data architectures: Demonstrate familiarity with UML, ERD, and custom architecture diagrams, including layered and microservices designs. Explain why architecture modelling requires various lenses to minimise the risk of miscommunication and help governance by incorporating data catalogues and dictionaries, ensuring robust metadata management	<p>This exciting one-month module exposes our Apprentices to the rich tapestry of key Data concepts. As the first stage of the ‘spiral curriculum’, we provide the learners with an inspiring panorama of what Data Engineering is really about. Deep-dives are provided in later modules, as we first model a real-world use case scenario that will be revisited throughout the course by gradually enriching it with detail.</p> <p>You are taught how the business generates value from data and the role played by Data Engineering in driving that function. You will learn fundamental Data Engineering best practices and standards, mastering essential skills for managing data projects and products, and understanding how to work with diverse stakeholders who use various types of data in a controlled way.</p> <p>Learners will work practically with modern data ecosystems and begin building their own toolbox of Big Data skills. Gaining insight into cutting-edge data architectures, from the very beginning of your programme, you will be introduced to governance and regulatory issues, data risks and key environmental considerations.</p>

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Recognise** fundamental data types and data source types, key Data Engineering technological standards and best practices, and relevant regulations.
 - K9, K10, K11, K16, K19, K21, K28
 - S13, S29,
 - B6
- **Apply** fundamental principles for fostering a data-driven culture through collaborating with diverse stakeholders in Agile and Lean teams, avoiding waste and respecting organisational processes for data quality and data project management.

- K4, K5, K7, K29, K30,
- S12, S22, S26, S27,
- B1, B2, B4, B6

- **Model** standardised Big Data ecosystems and architectures for enterprise data using visual approaches, **demonstrate the use of an approach** to designing a rudimentary data product, and **explain** how data products add value to the organisation.
 - K20, K24, K25, K26, K27,
 - S1, S2, S24,
 - B6

- **Explain** the importance of assessing, improving, and maintaining data quality using frameworks and methodologies to ensure data accuracy, completeness, and consistency.
 - K3, K4, K5, K26
 - S6, S26

Recommended Reading:

Haskell, C. (2024). Driving Data Projects: A comprehensive guide. British Computer Society [20 hours]

King, T., Schwarzenbach, J., (2020). Managing Data Quality: A practical guide. British Computer Society [15 hours]

Reis, J., Housley, M. (2022), Fundamentals of Data Engineering: Plan and Build Robust Data Systems, O'Reilly [25 hours]

Please see the core reading list at the end of this Programme Handbook.

Databases and Data Lakes

Module	Phase	Weekly Topics	Description
Data-bases and Data Lakes	Month 2-3	<p>Storing and Exploring data: Explore why businesses prioritize data retention through strategic Big Data utilisation. Delve into the differences and applications of Data Lakes, Data Warehouses, and Operational Data Stores, and learn data structuring concepts such as Schema-on-Write and Relational Modeling. Compare on-premise versus cloud storage options with examples. The webinar reinforces these concepts through interactive discussions, polls, and practical demonstrations in file system management and secure data storage with PowerShell. It also covers Data Lakehouse Architecture and deepens knowledge in Relational Modeling. Participants will develop SQL skills through hands-on DDL and DML exercises using SQL on AWS. Post-webinar content introduces a common semantic layer, covers business requirements for scalable databases, engages in AWS consolidation exercises, and explores network storage protocols and Linux filesystems, providing a comprehensive understanding of storage solutions including lakehouses, key-value stores, and columnar storage alongside secure storage strategies.</p> <p>Schemas and integration: In the evolving landscape of data management, the role of schemas and their integration into business systems is critical. This module delves into the advanced concepts of data modelling, exploring the intricacies of keys and relationships, schema evolution, and its profound impact on augmented analytics. Learners will gain a comprehensive understanding of data warehouses, data marts, and the distinctions between OLAP (Online Analytical Processing) and OLTP (Online Transaction Processing) systems. This topic also covers star schemas, data integration techniques, the design of SQL databases, metastores, and the use of open table formats such as Apache Iceberg and the integration frameworks like AWS Glue. Learners start practicing their SQL queries and normalization is introduced.</p> <p>Advanced SQL: In the data-driven enterprise, explore the impacts of upstream data structuring on downstream application, the ACID acronym and the importance of transactions in business use cases, writing nested CRUD SQL queries, performing joins and carrying out database normalisation, applying groupBy, aggregation and windowing. Appraising SQL database risks and disaster</p>	<p>This module explains the importance of structuring of upstream data for downstream consumption, while making sure that data does not get lost or corrupted.</p> <p>Apprentices will learn about how data is stored and queried using SQL and NoSQL, while being exposed to key database types, storage file formats and technical processes for data integration. We introduce key databases concepts like data normalisation and query optimisation, and the fundamental theory behind data consistency and availability to ensure that data is updated and accessed in guaranteed and reliable ways.</p> <p>You will learn skills of practically managing and monitoring the performance databases and data lake components, comprehend the significance of data quality, common database-related issues and their impact on data analysis. We focus on open-source standards throughout. Apprentices will be introduced to database performance monitoring.</p> <p>Data quality and performance EPA preparation receives emphasis in this module. You will also understand what regulatory considerations apply and how to keep your environment secure.</p>

		<p>recovery, portfolio work is introduced, focus: data quality and performance</p> <p>Database administration and optimisation: monitoring database operations, explaining database and query performance, and the security and regulatory considerations, listing relevant administration-related metrics, explaining operations cost and sustainability, explaining processing and storage costs involved parallelism, communicating downtime, optimising SQL queries, explaining partitioning and shards, project work is introduced, behaviours: timeliness and accountability</p> <p>NoSQL fundamentals: schema-on-read, BASE, CAP theorem, NRW, CAS, BSON and Parquet, MongoDB Atlas, NewSQL, overview of project deliverables including the presentation</p>	
--	--	---	--

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Apply** standard industry tools and best practices for designing, maintaining and optimising the performance and security of databases, data lakes, warehouses and lakehouses, including identifying relevant open-source standards for storing data.
 - K3, K9, K11, K14, K20,
 - S3, S4, S6, S12, S14, S24, S25
 - B5, B6
- **Explain** the fundamental concepts of SQL and NoSQL approaches to storing, updating and querying data and the impacts of changes in upstream database systems on the needs of diverse data consumers, and **list** the processing and storage costs involved.
 - K2, K12, K17, K19, K24, K25, K26
 - S6, S9, S12, S23, S24
 - B2
- **Evidence** the use of schemas, metadata, data modelling, transactions and identifying, monitoring and optimising key data access metrics to guarantee the reliability and sustainability of data and mitigate data risks.
 - K1, K5, K7, K15, K24, K25
 - S3, S5, S6, S7, S10, S13, S27
 - B4
- **Appraise** user and business requirements to **organise** upstream data for accuracy, completeness, consistency, timeliness and accessibility to downstream consumers whilst **acting proactively** taking accountability to ensure deadlines are met.
 - K4, K9
 - S1, S2, S5, S13, S26, B1

Recommended Reading:

Campbell, L., Majors, C. (2017). Database Reliability Engineering, O'Reilly Media [20 hours]

Done, P., Kamsky, A. (2023). Practical MongoDB Aggregations. Packt Publishing [25 hours]

King, T., Schwarzenbach, J., (2020). Managing Data Quality: A practical guide. British Computer Society [15 hours]

Perkins, L., Redmond, E., Wilson, J. (2018) Seven Databases in Seven Weeks. O'Reilly [25 hours]

Simsion, G., Witt, G. (2005) Data Modeling Essentials 3rd ed., Elsevier [20 hours]

Zhao, A. (2021). SQL Pocket Guide: A Guide to SQL Usage. O'Reilly Media [20 hours]

Programming and Scripting Essentials

Module	Phase	Weekly Topics	Description
Programming and Scripting Essentials	Months 3 -4	<ul style="list-style-type: none"> • Linux for Data Engineers: Linux use cases for Big Data, system structure, the filesystem, key utilities, job scheduling, scripting, grep, logs • Version control: DevOps basics for Data Products, intro to CI/CD, git, branching, merging, pull requests, GitHub, Conducting a code review • Python for Data Engineers: software development basics, syntax, coding conventions, control flow, functions, recursion, test-driven development • Data structures and OOP: collections, data serialisation, debugging, OOP concepts, design patterns • Data manipulation: working with APIs, the Pandas library, Pandas DataFrames, common errors, error handling and logging, other key libraries, regex • Algorithmic thinking: searching, sorting, graph algorithms, machine learning algorithms, algorithmic complexity, developing and validating algorithms, documenting algorithms • Parallel programming: concurrency, parallelism, distributed computing, Databricks, pyspark, RDD algorithms • Spark for Data Engineers: Data Pipelines essentials, SparkSQL, Spark Streaming, Spark clusters • Practical programming skills: Docker, data visualisation, writing a suite of tests, integration testing, security considerations, consolidation of the material 	<p>This module is designed to make Apprentices comfortable with programming, scripting and the software development practices in the modern data-oriented enterprise. Learners will script, code and conduct practical code reviews. There is an emphasis on practicing skills throughout this module and you will be applying the skills acquired to a practical work-based scenario, to develop a portfolio of evidence for the professional discussion element of the endpoint assessment.</p> <p>Focusing on Big Data programming, you will learn Linux, job scheduling, version control, Git, Python, data manipulation with pandas, parallel programming with Spark, DataBricks, algorithmic thinking, data visualisation, Docker, regular expressions, unit testing, the essentials of data pipelines, documenting, error handling and logging.</p> <p>This module contains one full-day Hackathon (on one of the regular teaching days).</p>

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Employ** software development tools and techniques for designing, deploying and maintaining secure data products and pipelines, including debugging, version control and testing.
 - K3, K6, K17, K20, K26
 - S9, S12, S13, S14, S17, S20, S25, S27
 - B1, B5, B6
- **Construct** algorithms that correctly and efficiently handle data at scale whilst mitigating risks.
 - K2, K5, K24, K25, K27
 - S1, S2, S3, S5, S16, S17, S24, S26
 - B2, B4
- **Demonstrate** the knowledge of the steps needed to prepare the code for production.
 - K6, K8, K28
 - S2, S4, S8, S14, S17, S20, S26
 - B1, B3

Networks Essentials and Cyber Security

Module	Phase	Topics	Description
Networks Essentials and Cyber Security	Month 5	<p>Introduction to Networks for Data Engineers:</p> <ol style="list-style-type: none"> 1. Explain the principles of computer networks, including IP addressing, TCP/IP and the OSI model. 2. Demonstrate awareness of modern networking practices 3. Learn about network infrastructure costs and sustainability. 4. Explain the role of different types of network devices <p>Cyber Security essentials:</p> <ol style="list-style-type: none"> 1. Understand the fundamental principles of the CIA triad and its application in cyber security. 2. Recognise risks, vulnerabilities, and threats to ensure robust security for data products. 3. Explain security controls and quantify and evaluate the impact of security breaches 4. Identify and mitigate common cyber threats <p>Designing secure architectures:</p> <ol style="list-style-type: none"> 1. Implement functionality of data products and data pipelines in a way that protects data at rest and data in transit using a zero-trust approach 2. Make use of advanced security techniques to safeguard data products and data pipelines against unauthorised access and security breaches. 3. Practise visual methods of designing secure architectures through diagrams and flowcharts to improve data product and data pipeline communications. <p>Security policy and incident response:</p> <ol style="list-style-type: none"> 1. Explain the importance of security policies and established policy frameworks in keeping data products and pipelines secure 2. Implement risk mitigation strategies and remediation techniques to address potential security breaches. 3. Document and learn from security incidents to enhance future response strategies. 4. Identify various types of security policies and their applications to enhance organisational security. 	<p>This module is designed to make Apprentices comfortable with network infrastructures and the cyber security risks associated with networked environments. This is not a generic networks/security module, but rather, it has been tailored to the needs of the Level 5 Data Engineer apprenticeship standard.</p> <p>Networks Essentials and Cyber Security introduces Level 5 Data Engineer learners to essential networking and security concepts that are necessary to set the scene before the learners can comfortably study further modules (based around building secure distributed data products that use cloud set-ups as well as on-premise clusters and data centres). Apprentices will learn about the principles of networks and cyber security, which will prepare them for designing data ingestion and data pipelines. Having learned about sustainable network infrastructures, vulnerability types and worked examples of evaluating impact, they will gain a comprehensive insight into cyber threats, their mitigation and remediation.</p> <p>Learners will practice their skills of working in accordance with security policies, responding to incidents and sharing lessons learnt. They will acquire skills for layering security controls and enabling extra security, managing security requirements, designing secure networked architectures such as clusters, while identifying threats and opportunities. The knowledge gained in this module includes TCP/IP networking, the OSI model, subnetting, the CIA and AAA triad concepts, encryption, secure systems modelling, infrastructure costs and working with real-world security frameworks.</p>

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Apply** key networking concepts and tools when designing, maintaining, costing and documenting secure data architectures, including **preparing** secure data pipelines and secure data products.
 - K2, K12, K13, K14, K20, K26
 - S1, S3, S5, S12, S14, S25, S27
 - B1, B6
- **Evaluate** networking and cyber security threats, opportunities, technical gaps and lessons learned from investigations, including **assessing** the state of encryption and access control.
 - K20, K25, K26, K23, K28, K30
 - S8, S10, S12, S14, S20, S24, S25
 - B2, B5
- **Demonstrate** knowledge of cyber security fundamental concepts, key types of security policies policy-driven incident response types and relevant regulatory requirements.
 - K1, K9, K10, K11, K21, K22, K25, K26, K30
 - S5, S8, S13, S18, S19
 - B1, B6

Recommended Reading:

Chou, E. (2020). Mastering Python Networking. Packt Publishing [20 hours]

Cloud Engineering

Module	Phase	Weekly Topics	Description
Cloud Engineering	Months 6-7	<p>Cloud fundamentals:</p> <p>Introduction to cloud compute types and costs; the business case for cloud and calculating ROI on cloud projects; Netflix case study; overview of key business and user requirements for Cloud-based data products, the role of hyperscalers in edge computing, overview of typical categories of Cloud utilities and resources across top 3 providers, a walk-through explaining key IAM concepts and the importance of using Service Accounts for main types of use cases</p> <p>Data in the cloud:</p> <p>How do businesses collate and review their requirements for cloud data movement and storage, estimating data egress costs for different data groups (columnar, key-value, relational and blob storage); regulatory risks for storing and moving data in the cloud; pros and cons of Shared Key/Active Directory/ generating and using a shared access signatures(SAS) for cloud data access control, pros and cons of the CQRS pattern in writing data to the cloud (example walk-through: event sourcing with Azure Cosmos DB)</p> <p>Containers and orchestration:</p> <p>The key skill of visual modelling for container design (practice drawing simple diagrams), comparing the pros and cons of serverless mode vs. the VM deployment model, practical example of using serverless mode for deployments, a walk-through of deploying a simple Python microservices application using Docker, Kubernetes and serverless deployments, a walk-through of reporting on resource usage and optimizing resources for sustainable growth</p> <p>Securely deploying cloud data products:</p> <p>an overview of CI/CD in the cloud, security risks of cloud deployments, a walk-through of Azure DevOps, cloud computing execution models - list different ones , how to securely used cloud-based APIs, how to</p>	<p>In this module, you will advance your knowledge of Cloud fundamentals, managing cloud projects and designing and maintaining cloud data products using microservices, as well as advanced Cloud topics in Containers and Orchestration using core Azure services. At every step, you will be introduced to cloud-specific regulatory risks to ensure that you work professionally in a risk-aware manner.</p> <p>Importantly, you will not just learn Azure, instead you will be shown how your knowledge of the cloud can be applied across main cloud providers such as Google cloud and AWS. Towards this end, we will focus on transferable skills.</p> <p>For most exercises, you will be shown how to create a free student Azure account to perform your tasks.</p> <p>Lots of practical work here focusing on cloud deployments (including serverless deployments), testing and solutions integration with popular cloud services such as Snowflake.</p> <p>The module culminates in the release of a Cloud-native data product dashboard designed and deployed by the Apprentice. Apprentices will be asked to reflect on recommending potential opportunities for use in own department and develop collaborative relationships with stakeholders at work.</p> <p>This module contains one full-day Hackathon (on one</p>

		<p>securely use scalable cloud-based analytics products and lakehouses, security considerations of cloud data integration (walk-through of Data Factory), Site Reliability Engineering (SRE) principles; walk-through: monitoring the availability and downtime using dashboards, a walkthrough of working with Cloud logs and spotting and remediating problems,</p> <p>Migrations, Archiving, Monitoring and Disaster Recovery:</p> <p>A selection of narrated case studies of different types of cloud migration projects including failed ones and how to document lessons learned, the importance of robust Archiving policies and frameworks; the significance of introducing redundancy to the data for performance and availability reasons including focus on best practices, focus on DR policies including incident response, a walk-through of Azure Monitor console or similar, a walk-through of writing and deploying a cloud-native monitoring service in Python</p> <p>Cloud solutions integration:</p> <p>the importance of integration testing, integrating different cloud providers into a hybrid cloud; integrating on-premise clusters with cloud; case study: DevTest Labs, a recap of the significance of open-source standards; an introduction to Snowflake – case study and walkthrough of integration, recap of Databricks – case study and walkthrough of integration; extend: an overview of PowerBI and PowerQuery and how they can be integrated with modern hybrid clouds</p> <p>Putting it all together:</p> <p>what do we mean by full-stack cloud solution designs; a focus on when cloud is definitely not needed; sanity checking user and business requirements, walk-through: designing and deploying a Python data product dashboard in the cloud that links to visualization (consider free Tableau account for students), monitoring the set-up cost and running cost of complex multi-component integrated data products that are cloud-native, net-zero: right-sizing</p>	<p>of the regular teaching days).</p>
--	--	--	---------------------------------------

		instances, implementing auto-scaling policies;	
--	--	--	--

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Apply** key cloud concepts and tools when designing, costing, deploying, maintaining, troubleshooting and documenting secure cloud-native data products.
 - K6, K11, K12, K14, K15, K17, K20, K24, K25, K26
 - S1, S2, S3, S7, S9, S12, S13, S18, S22, S27
 - B1, B3, B4, B5, B6
- **Evaluate** business rationale, sustainability and regulatory risks of cloud-based deployments and **monitor** the ongoing cost, availability and service health of cloud-native data products.
 - K1, K4, K7, K9, K10, K12, K13, K14, K17, K22
 - S8, S11, S12, S18, S19, S22, S24, S25, S26, S27, S28
 - B1, B3, B4, B5, B6
- **Demonstrate** practical knowledge of data movement and storage, working with cloud-native APIs and the principles of site reliability engineering (SRE).
 - K2, K7, K8, K13, K14, K22, K24
 - S9, S14, S18, S19
 - B1, B2, B3
- **Demonstrate** the knowledge of the steps needed to prepare the code for deployment in the cloud, focusing on CI/CD, containers, orchestration and serverless deployment.
 - K6, K8, K28
 - S2, S4, S8, S14, S17, S20, S26
 - B1, B3
- **Apply** cloud-native frameworks to store data, and to integrate it with non-cloud sources, or with other clouds.
 - K1, K5, K8, K16, K17, K21
 - S1, S2, S3, S7, S9, S15, S24, S25
 - B1, B2, B5, B6
- **Evidence** the awareness of sustainability considerations such as auto-scaling and right-sizing instances to promote net-zero objectives.
 - K1, K5, K7, K15, K24, K25
 - S3, S5, S6, S7, S10, S13, S27
 - B4

Data Collection and Ingestion pt. 1

Module	Phase	Weekly Topics	Description
Data Collection and Ingestion pt. 1	Months 7 - 8	<ul style="list-style-type: none"> Introduction to collecting data: <ol style="list-style-type: none"> Justify the importance of automation in data collection and ingestion Evaluate common data cleaning techniques Recognise the steps required to pre-process data for machine learning purposes Demonstrate practical data collection skills Heterogenous data ingestion patterns: <ol style="list-style-type: none"> Explain the key design and architecture considerations of heterogeneous data ingestion systems Report on the usefulness of heterogeneous ingestion patterns for business use cases Demonstrate practical application of file ingestion combining multiple sources of data APIs and microservices for Data Engineers: <ol style="list-style-type: none"> Explain the benefits of microservices architectures in data engineering Demonstrate ability to use an API to ingest data Design a data ingestion architecture using APIs and microservices Data collection quality and ingestion risks: <ol style="list-style-type: none"> Explain the benefits of setting up SLAs around collecting and ingesting data Explain how you would manage and report on data risks in a business environment Apply strategies for collecting and ingesting PII and sensitive data Evidence data quality improvement practices in practical scenarios Compress ingested data in Python to reduce sustainability risks and support net-zero goals 	<p>Collecting rich data is a skill needed in the modern workplace and different approaches are introduced in this module.</p> <p>You will acquire the skills needed to plan data acquisition and evaluate data quality as it enters the enterprise, considering relevant validation checks, metrics and risk measures.</p> <p>There is a focus on good architectural practices, with data staging areas, APIs and microservices receiving special attention. Main architectural patterns are presented.</p> <p>Learners will be introduced to the essentials of data pipelines, ETL and ELT, the topics which will be explored in further modules, noting how data collection and ingestion relates to the early stage of data pipelines and ETL/ELT.</p>

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- Demonstrate** practical knowledge of data collection and ingestion and its position within organisational ETL/ELT strategies, **deciding** relevant validation checks, metrics and risk measures, with a focus on data quality and availability.
 - K2, K7, K8, K14, K18, K19, K20, K22, K25, K30
 - S1, S9, S14, S15, S18, S19, S25, S26, S27
 - B1, B4, B5, B6
- Model** standardised data ingestion architectures and **explain** how to design microservices-oriented data products, and how API and microservices add value to the data infrastructure.
 - K20, K22, K24, K25, K26

- S1, S2, S12, S13, S14, S15, S24, S25
- B2, B4, B5, B6
- **Design** hybrid data collection and ingestion pipelines containing cloud and on-premise elements.
 - K20, K22, K24, K25, K26
 - S1, S2, S12, S13, S14, S15, S24, S25
 - B2, B4, B5, B6
- **Apply** data collection and ingestion best practices to select and stage external data, and prepare and enrich it with necessary and optional attributes to integrate it in a way that adds value.
 - K1, K5, K8, K16, K17, K21
 - S1, S2, S3, S7, S9, S15, S24, S25
 - B1, B2, B5, B6
- **Evidence** the use of sustainability techniques such as feature engineering and storage and traffic compression for designing and optimising data ingestion and collection solutions, to promote net-zero goals and mitigate data risks.
 - K1, K5, K7, K15, K24, K25
 - S3, S5, S6, S7, S10, S13, S27
 - B4

Recommended Reading:

Espenschutz, G. (2023). Data Ingestion with Python – Cookbook. Packt Publishing [30 hours]

King, T., Schwarzenbach, J., (2020). Managing Data Quality: A practical guide. British Computer Society [15 hours]

Mertz, D. (2021) Cleaning Data for Effective Data Science. Packt Publishing. [20 hours]

Reis, J., Housley, M. (2022), Fundamentals of Data Engineering: Plan and Build Robust Data Systems, O'Reilly [25 hours]

Data Collection and Ingestion pt. 2

Module	Phase	Weekly Topics	Description
Data Collection and Ingestion pt. 2	Month 9	<p>Event-driven architectures: EDA fundamentals, pub-sub model, messages, topics, event streaming with Kafka, consumer groups, partitions, clusters</p> <p>Advanced Kafka scenario: Azure Event Hubs vs Event Grid, SASL, auto-inflate, troubleshooting skills, root-cause analysis</p> <p>Python libraries for rich data collection: kafka-python, CRC32, Avro, schema registry interaction, Rekognition, Textract, OpenCV, LibROSA, Scrapy</p> <p>Monitoring an ingestion service: automation, simple forecasting, anomaly detection, integration with incident management</p>	<p>In this module, you'll be walked through the requirements for the End-Point Assessment (EPA) during the scenario demonstration that includes innovative learning techniques, root-cause analysis and a real-world troubleshooting case study.</p> <p>You will participate in an open scenario practice, which has been meticulously designed to encompass all the knowledge and skills acquired in the course up to this point.</p> <p>Furthermore, the module will focus on group work, encouraging you to collaboratively work through potential solutions and present your scenario as a team.</p>

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Demonstrate** practical knowledge of event-driven architectures and rich data collection.
 - K2, K7, K8, K14, K18, K19, K20, K22, K25, K30
 - S1, S9, S14, S15, S18, S19, S25, S26, S27
 - B1, B4, B5, B6
- **Analyse** data collection and ingestion issues, identifying and remediating root causes.
 - K20, K22, K24, K25, K26
 - S1, S2, S12, S13, S14, S15, S24, S25
 - B2, B4, B5, B6
- **Apply** data engineering best practices to automate, monitor and optimise data collection and ingestion.
 - K1, K5, K8, K16, K17, K21
 - S1, S2, S3, S7, S9, S15, S24, S25
 - B1, B2, B5, B6

Recommended Reading:

Espenschutz, G. (2023). Data Ingestion with Python – Cookbook. Packt Publishing [30 hours]

Data Pipelines

Module	Phase	Topics	Description
Data Pipelines	Months 9-10	<p><i>Data Integration Techniques Beyond ETL</i></p> <ul style="list-style-type: none"> • Advanced Data Integration: <ul style="list-style-type: none"> ○ Techniques for integrating data from diverse sources ○ Real-time data integration and streaming ○ Data blending and data federation • Tools and Technologies: <ul style="list-style-type: none"> ○ Overview of modern data integration tools (e.g., Apache NiFi, Talend) ○ Case studies of advanced data integration implementations <p><i>2. Workflow Management for Data Pipelines</i></p> <ul style="list-style-type: none"> • Designing Efficient Workflows: <ul style="list-style-type: none"> ○ Principles of workflow management in data engineering ○ Creating and managing data pipeline workflows ○ Tools for workflow orchestration (e.g., Apache Airflow, Prefect) • Optimisation and Automation: <ul style="list-style-type: none"> ○ Techniques for optimising data workflows ○ Automation strategies for repetitive tasks ○ Monitoring and maintaining workflow efficiency <p><i>3. On-Premise Resource Management</i></p> <ul style="list-style-type: none"> • Resource Allocation: 	<p>The Data Pipelines module provides candidates with advanced skills and knowledge in data integration, workflow management, resource management, scalability, and service management. It also includes practical hackathons focused on data cleansing, enrichment, testing, and debugging to ensure hands-on experience.</p> <p>Using Cloud and non-cloud (Airflow) environments, the learner is exposed to workflow management for getting enterprise data to flow from A to B.</p> <p>Practical skills include data cleansing and enrichment, testing and debugging, cluster resource management and service-level agreements around data.</p> <p>Apprentices are encouraged to seek out in-house data pipeline use cases to collaborate on at their workplace, while meeting regularly with their coach to ensure that all their on-the-job training is logged.</p> <p>This module contains one full-day Hackathon (on one of the regular teaching days).</p>

		<ul style="list-style-type: none">○ Managing computational resources for data pipelines○ Techniques for efficient resource utilisation○ Load balancing and resource scheduling● Infrastructure Management:<ul style="list-style-type: none">○ Setting up and maintaining on-premise data infrastructure○ Tools for infrastructure management (e.g., Kubernetes, Docker)○ Case studies of on-premise resource management	
		<p><i>4. Scalability - From Data Pipelines to Data Architectures</i></p> <ul style="list-style-type: none">● Scaling Data Pipelines:<ul style="list-style-type: none">○ Strategies for scaling data pipelines to handle larger volumes of data○ Ensuring performance and reliability at scale○ Tools and techniques for scalable data processing (e.g., Apache Spark, Kafka)● Architectural Considerations:<ul style="list-style-type: none">○ Designing data architectures for scalability and flexibility○ Balancing performance with cost and resource constraints○ Case studies on scalable data architectures <p><i>5. Service Management and EPA Preparation</i></p> <ul style="list-style-type: none">● Service Management:<ul style="list-style-type: none">○ Principles of effective service management in data engineering	

		<ul style="list-style-type: none"> ○ Monitoring and maintaining data pipeline services ○ Incident management and response • EPA Preparation: <ul style="list-style-type: none"> ○ Key concepts and skills required for EPA (End Point Assessment) ○ Strategies for effective preparation and practice ○ Reviewing past EPA examples and best practices <p><i>6. Practical Data Cleansing and Enrichment (Hackathon)</i></p> <ul style="list-style-type: none"> • Data Cleansing Techniques: <ul style="list-style-type: none"> ○ Identifying and correcting data quality issues ○ Standardisation, deduplication, and validation techniques ○ Tools for data cleansing (e.g., OpenRefine, Trifacta) • Data Enrichment: <ul style="list-style-type: none"> ○ Techniques for enhancing data with additional information ○ Integrating external data sources for enrichment ○ Practical exercises in data cleansing and enrichment <p><i>7. Testing and Debugging Data Pipelines (Hackathon)</i></p> <ul style="list-style-type: none"> • Testing Strategies: <ul style="list-style-type: none"> ○ Techniques for testing data pipelines and ensuring reliability ○ Automated testing frameworks and tools ○ Best practices for test coverage and test case design • Debugging Techniques: 	
--	--	---	--

		<ul style="list-style-type: none"> ○ Identifying and resolving common issues in data pipelines ○ Tools and methods for effective debugging ○ Hands-on exercises in testing and debugging 	
--	--	---	--

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Demonstrate practical knowledge of workflow management and ETL/ELT.**
 - K2, K7, K8, K14, K18, K19, K20, K22, K25, K30
 - S1, S9, S14, S15, S18, S19, S25, S26, S27
 - B1, B4, B5, B6
- **Analyse data availability and quality issues**, identifying and remediating root causes.
 - K20, K22, K24, K25, K26
 - S1, S2, S12, S13, S14, S15, S24, S25
 - B2, B4, B5, B6
- **Prepare data for analysis by cleansing and enrichment.**
 - K1, K5, K8, K16, K17, K21
 - S1, S2, S3, S7, S9, S15, S24, S25
 - B1, B2, B5, B6

Data Responsibilities

Module	Phase	Topics	Description
Data Responsibilities	Month 11 - 12	<p><i>1. Legal Aspects of Data Engineering and Privacy-by-Design</i></p> <ul style="list-style-type: none"> • Legal Frameworks: <ul style="list-style-type: none"> ○ Overview of key data protection laws (e.g., GDPR, DPA) ○ Understanding the legal obligations of data engineers ○ Case studies on legal compliance in data engineering • Privacy-by-Design: <ul style="list-style-type: none"> ○ Principles of privacy-by-design and their application ○ Techniques for embedding privacy into data systems ○ Tools and frameworks to support privacy-by-design <p><i>2. Ethics and Anonymisation</i></p> <ul style="list-style-type: none"> • Data Ethics: <ul style="list-style-type: none"> ○ Understanding ethical considerations in data engineering ○ Balancing innovation with ethical responsibilities ○ Real-world examples of ethical dilemmas and resolutions • Anonymisation Techniques: <ul style="list-style-type: none"> ○ Methods for anonymising data while preserving utility ○ Legal and ethical implications of data anonymisation ○ Tools and technologies for data anonymisation <p><i>3. Data Governance and Stewardship</i></p> <ul style="list-style-type: none"> • Data Governance: <ul style="list-style-type: none"> ○ Establishing a data governance framework ○ Roles and responsibilities in data governance ○ Policies and procedures for data management • Data Stewardship: <ul style="list-style-type: none"> ○ Principles of data stewardship ○ Ensuring data quality and integrity ○ Strategies for effective data stewardship <p><i>4. Compliance and Reporting on Risks</i></p> <ul style="list-style-type: none"> • Compliance Requirements: <ul style="list-style-type: none"> ○ Identifying relevant compliance standards and regulations 	<p>The Data Responsibilities module equips candidates with the knowledge and skills to navigate the legal, ethical, and governance aspects of data engineering. It focuses on privacy-by-design, data governance, compliance, risk management, and effective stakeholder engagement.</p> <p>Roles and responsibilities in a data-driven enterprise are introduced with a focus on data governance, compliance, privacy and security.</p> <p>The concepts of Data stewardship and leadership are elaborated upon, and the learners will master the skills of stakeholder engagement and communications including presentation skills for their endpoint assessment project component.</p> <p>A group activity/project is recommended in this module.</p>

		<ul style="list-style-type: none"> ○ Implementing compliance strategies and best practices ○ Tools for monitoring and ensuring compliance ● Risk Reporting: <ul style="list-style-type: none"> ○ Techniques for assessing and reporting data-related risks ○ Creating comprehensive risk reports ○ Communicating risks to stakeholders <p><i>5. Stakeholder Engagement and Communication</i></p> <ul style="list-style-type: none"> ● Effective Communication: <ul style="list-style-type: none"> ○ Strategies for engaging with stakeholders ○ Techniques for clear and effective communication ○ Building and maintaining stakeholder relationships ● Team Roles and Responsibilities: <ul style="list-style-type: none"> ○ Understanding the roles within data teams ○ Defining and allocating responsibilities ○ Collaboration and consolidation within data teams <p>Content covered:</p> <ul style="list-style-type: none"> ● Data Governance frameworks ● Data stewardship and leadership <ul style="list-style-type: none"> ● The lifecycle of data ● Roles and responsibilities in a data-driven enterprise <ul style="list-style-type: none"> ● Data compliance tools ● Data Privacy and Security ● Privacy by design and data anonymisation techniques ● Stakeholder engagement and communications skills <ul style="list-style-type: none"> ● Reporting on risks 	
--	--	---	--

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Demonstrate** awareness of data governance practices and frameworks and the roles and responsibilities in a data-driven enterprise.
 - K3 K4 K6 K10 K11
 - S5 S13 S17
 - B1, B2, B3, B4, B5, B6
- **Explain and demonstrate the usefulness of** data stewardship and applying best practices in maintaining data privacy and security.
 - K9, K10, K11, K12
 - S17, S22, S29
 - B1, B2, B3, B4, B5, B6
- **Explain and demonstrate the usefulness of** stakeholder engagement and communication, and report on data risks as appropriate.
 - K5, K30
 - S19, S23
 - B1, B2, B3, B4, B5, B6

Production Environment

Module	Phase	Weekly Topics	Description
Production Environment	Months 13 - 14	<p><i>1. Advanced Testing and Deployment Strategies</i></p> <ul style="list-style-type: none"> Advanced Testing: <ul style="list-style-type: none"> User Acceptance Testing (UAT) <ul style="list-style-type: none"> UAT processes and stakeholder involvement Creating and managing UAT plans Stress Testing and Performance Tuning <ul style="list-style-type: none"> Tools and techniques for stress testing (e.g., JMeter, Gatling) Analysing results from stress tests Performance tuning and optimisation based on test results Deployment Strategies: <ul style="list-style-type: none"> Blue-Green Deployments and Canary Releases Zero-Downtime Deployments Implementing Rollbacks and Roll-forwards Automating Deployment with Configuration Management Tools (e.g., Ansible, Chef) Rsync, ssh, github actions, etc <p><i>2. MLOps and Model Deployment</i></p> <ul style="list-style-type: none"> MLOps Principles: <ul style="list-style-type: none"> Overview of MLOps and its importance in production 	<p>This module prepares the Apprentice to work in different production cultures such as Secure DevOps and MLOps, whilst teaching practical skills of continuous Integration, Integration testing and logging, Visualising performance, maintenance, documentation, deployment strategies, UAT, stress testing and service management.</p> <p>Team exercises aim to boost your confidence and competence while reinforcing the ability to work through potential solutions and present them as a group.</p> <p>Additional Notes:</p> <ul style="list-style-type: none"> Team Exercises and Collaboration: <ul style="list-style-type: none"> Emphasise hands-on exercises and team-based projects. Real-world scenarios to apply learned skills. Presentation of solutions and peer reviews to build confidence and competence.

		<ul style="list-style-type: none">○ Integration of ML workflows into CI/CD pipelines○ Tools and frameworks for MLOps (e.g., MLflow, Kubeflow)• Model Deployment:<ul style="list-style-type: none">○ Techniques for deploying ML models (e.g., Docker, Kubernetes)○ Monitoring and logging ML models in production○ Handling model versioning and rollback strategies• Model Management:<ul style="list-style-type: none">○ Continuous training and updating models in production○ Ensuring data quality and consistency for retraining○ Case studies and best practices for MLOps in production <p><i>3. Technical Debt Management and Software Maintenance</i></p> <ul style="list-style-type: none">• Managing Technical Debt:<ul style="list-style-type: none">○ Identifying and quantifying technical debt○ Strategies for reducing and managing technical debt○ Prioritisation techniques for technical debt repayment○ Balancing new features with technical debt management• Software Maintenance:<ul style="list-style-type: none">○ Best practices for ongoing software maintenance○ Version control for maintaining codebases○ Documenting changes and maintaining comprehensive logs	
--	--	--	--

		<ul style="list-style-type: none">○ Change management processes and tools (e.g., ITIL Framework) <p><i>4. Performance Monitoring and Service Management</i></p> <ul style="list-style-type: none">● Performance Monitoring:<ul style="list-style-type: none">○ Advanced metrics and KPIs for production systems○ Setting up and customising monitoring dashboards○ Tools for visualising performance metrics (e.g., Grafana, Kibana)○ Analysing performance trends and identifying bottlenecks● Service Management:<ul style="list-style-type: none">○ Incident management and response○ Incident response plans and playbooks○ Real-time monitoring and alerting○ Post-incident analysis and reporting○ Service-level agreements (SLAs) and root cause analysis <p><i>5. Change Management</i></p> <ul style="list-style-type: none">● Quality Assurance:<ul style="list-style-type: none">○ Automated testing frameworks and tools○ Continuous quality monitoring○ Integration testing and logging● Robustness and Scalability:<ul style="list-style-type: none">○ Designing for scalability and high availability○ Load balancing and resource optimisation	
--	--	--	--

		<ul style="list-style-type: none"> ○ Case studies on robust system designs • Documentation and Change Management: <ul style="list-style-type: none"> ○ Maintaining comprehensive and up-to-date documentation ○ Best practices for documenting changes ○ Tools and techniques for effective change management 	
--	--	--	--

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Recognise** technical debt and implement metrics to choose appropriate technologies and sunset legacy systems.
 - K14, K21, K25, K28, K29
 - S12, S21
 - B4, B6
- **Demonstrate** familiarity with DevOps, DevSecOps, MLOps, continuous integration and managing data services from the perspective of a modern data-driven production environment.
 - K21, K22, K25, K26
 - S10, S12, S13, S14, S17, S20
 - B2, B4, B5, B6
- **Demonstrate** knowledge of deployment processes, deployment best practices and automation.
 - K6, K8, K28
 - S4, S8, S14, S17, S20, S26
 - B1, B3

Data Products

Module	Phase	Weekly Topics	Description
Data Products	Months 14 - 15	<p><i>Advanced Data Products - Beyond Reports and Dashboards</i></p> <ul style="list-style-type: none"> • Overview of Advanced Data Products: <ul style="list-style-type: none"> ○ Predictive analytics models ○ Recommendation systems ○ Data-driven decision support systems ○ Data APIs and microservices for product integration • Practical Applications: <ul style="list-style-type: none"> ○ Industry-specific data product examples (e.g., healthcare, finance, retail) ○ Case studies on innovative data product implementations <p><i>2. User Requirements and Human-Centric Design</i></p> <ul style="list-style-type: none"> • Gathering User Requirements: <ul style="list-style-type: none"> ○ Techniques for eliciting detailed requirements from end-users ○ Conducting user interviews and workshops ○ Creating user personas and scenarios • Human-Centric Design: <ul style="list-style-type: none"> ○ Principles of user experience (UX) design for data products ○ Wireframing and prototyping data products ○ Iterative design processes and incorporating user feedback ○ Ensuring accessibility and inclusivity in data product design 	<p>This module equips candidates with the skills to develop, maintain, and improve a variety of data products, with a special focus on building AI-enabled data products. It covers understanding and integrating user and business requirements, creating compelling visualisations, ensuring product resilience, and leveraging AI and machine learning technologies.</p> <p>This module also prepares the candidate to collaborate with stakeholders. The practical skills include gathering requirements, troubleshooting common issues, data visualisation and dashboarding. There is an introduction to AI/Machine Learning, data analytics, reporting and evaluating the success of a data product.</p> <p>In this module, you'll be able to get a full understanding of how the portfolio is utilised during the professional discussion.</p> <p>You will participate in professional discussion practice, designed to replicate the End Point Assessment (EPA) experience as closely as possible. This exercise aims to boost your confidence and competence while reinforcing the ability to communicate your knowledge and skills as supported by your portfolio.</p>

		<p><i>3. Business Requirements - From Scalability to Compliance</i></p> <ul style="list-style-type: none">• Understanding Business Requirements:<ul style="list-style-type: none">○ Translating business objectives into data product features○ Aligning data products with organisational goals○ Prioritising features based on business impact• Scalability and Performance:<ul style="list-style-type: none">○ Designing data products for scalability and high performance○ Techniques for ensuring data product resilience and reliability○ Implementing fault tolerance and disaster recovery strategies• Compliance and Security:<ul style="list-style-type: none">○ Understanding regulatory requirements (e.g., GDPR, HIPAA)○ Ensuring data security and privacy in data product design○ Implementing compliance monitoring and reporting <p><i>4. Data Visualisation and Storytelling</i></p> <ul style="list-style-type: none">• Advanced Data Visualisation Techniques:<ul style="list-style-type: none">○ Principles of effective data visualisation○ Creating interactive visualisations and dashboards○ Tools and technologies for data visualisation (e.g., D3.js, Power BI, Tableau)• Telling a Story with Data:<ul style="list-style-type: none">○ Crafting compelling narratives using data○ Presenting data insights to non-technical audiences○ Techniques for effective data storytelling	
--	--	---	--

		<ul style="list-style-type: none"> ○ Case studies of successful data storytelling <p><i>5. Building an AI-Enabled Data Product</i></p> <ul style="list-style-type: none"> • Introduction to AI and Machine Learning: <ul style="list-style-type: none"> ○ Overview of AI and ML concepts ○ Identifying opportunities for AI in data products • Designing an AI-Enabled Data Product: <ul style="list-style-type: none"> ○ Integrating machine learning models into data products ○ Ensuring data quality for AI applications ○ Handling model training and deployment • Implementation and Evaluation: <ul style="list-style-type: none"> ○ Building and deploying an AI-enabled data product ○ Monitoring and maintaining AI models in production ○ Evaluating the performance and impact of AI features ○ Continuous improvement of AI models based on feedback and new data 	
--	--	---	--

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Explain** how to collaborate with Data Scientists and Data Analysts to build Data Products that conform to user and business requirements.
 - K25, K28, K30
 - S1, S2, S3, S5, S17, S22
 - B1, B2, B3, B5
- **Evaluate** the success of data products.
 - K4, K5, K7, K29, K30,
 - S12, S22, S26, S27,
 - B1, B2, B3, B4, B5, B6

Data Futures

Module	Phase	Weekly Topics	Description
Data Futures	Month 15	<p><i>Evaluating Technology Success and Managing Innovation</i></p> <ul style="list-style-type: none"> Evaluating Technology Success: <ul style="list-style-type: none"> Metrics for assessing the success of technology implementations Techniques for conducting post-implementation reviews Analysing technology performance and impact on business outcomes Managing Innovation: <ul style="list-style-type: none"> Strategies for fostering a culture of innovation Managing innovation projects and overcoming challenges Techniques for staying ahead of technological trends and advancements <p><i>2. Continuous Professional Development Directions</i></p> <ul style="list-style-type: none"> Career Development Strategies: <ul style="list-style-type: none"> Identifying opportunities for professional growth and advancement Developing a personal career development plan Networking and building professional relationships Skill Enhancement: <ul style="list-style-type: none"> Keeping up-to-date with emerging technologies and industry trends 	<p>The Data Futures module prepares candidates to evaluate the success of technology implementations and manage innovation effectively. It emphasises the importance of continuous professional development and revisits key sustainability considerations in technology.</p> <p>In the final weeks, candidates discuss Data Ethics and CPD directions, and evaluating the viability of innovative technological stacks through horizon scanning and scorecards.</p> <p>Apprentices will focus on Continuous improvement and professional growth for their EPA preparation.</p> <p>Apprentices are meeting regularly with their coach to ensure that all their on-the-job training is logged. Candidates will finalise the topics for their EPA projects before the gateway.</p>

		<ul style="list-style-type: none"> ○ Participating in professional development activities (e.g., workshops, conferences, online courses) ○ Leveraging certifications and further education to advance career prospects <p><i>3. Re-cap of Sustainability Considerations</i></p> <ul style="list-style-type: none"> • Sustainability in Technology: <ul style="list-style-type: none"> ○ Understanding the environmental impact of technology solutions ○ Best practices for designing sustainable technology systems ○ Incorporating sustainability into data management and storage practices • Corporate Social Responsibility (CSR): <ul style="list-style-type: none"> ○ The role of technology in supporting CSR initiatives ○ Strategies for aligning technology practices with corporate sustainability goals ○ Case studies of successful sustainability initiatives in the tech industry 	
--	--	---	--

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Discuss** Data Ethics and CPD directions, evaluating the viability of innovation in Data.
 - **K28 K29**
 - S12 S28 S29
 - B1, B2, B3, B4, B5, B6
- **Appraise** the significance of the human factors of sociotechnical systems and apply user research to improve the user experience of systems.
 - **K7, K24, K30**
 - S5, S11, S17, S23, S24
 - B3, B4, B5, B6

End Point Assessment (EPA) Gateway

At the end of the 15-month (standard duration) teaching and learning practical period, your manager must be content that you have attained sufficient KSBs to complete the apprenticeship. Your employer may take advice from BPP, but it is your employer's decision, not BPP's. You will then be able to enter the so-called "gateway".

You must meet the gateway requirements before starting your EPA:

- You will need to confirm you are ready to take the EPA
- You will need to have achieved the mandatory English and Mathematics standard in line with the apprenticeship funding rules

At gateway, you will:





submit a project brief for the project evaluation report, presentation and questions

confirm that you have completed a suitable project that meets the knowledge, skills and behaviours.

submit all the gateway evidence to the EPAO (end-point assessment organisation), including any organisation specific policies and procedures requested by the EPAO – your coach will guide you through the process

End Point Assessment (EPA)

It is **important** to read all the assessment guide documents, as they contain important details.

Assessment Guide Documents	Check box when you have read this document
<ul style="list-style-type: none"> Apprenticeship standards: https://www.instituteforapprenticeships.org/apprenticeship-standards/data-engineer-v1-0 	<input type="checkbox"/>
<ul style="list-style-type: none"> IfATE Assessment Plan: https://www.instituteforapprenticeships.org/apprenticeship-standards/data-engineer-v1-0?view=epa 	<input type="checkbox"/>
<ul style="list-style-type: none"> EPA guidance from BCS <ul style="list-style-type: none"> BCS End-point Assessment Guide  Data Engineer Apprenticeship Level 5 EPA Guide V1.0.pdf Assessment Method 1 – Guide  L5 Data Engineer IfATE V1.0 AM1 Apprentice and Employer Guide V1.0.pdf Assessment Method 1 – Checklist  L5%20Data%20Engineer%20IfATE%20V1.0 Assessment Method 1 – Sign-off Document  L5%20Data%20Engineer%20IfATE%20V1.0 	<input type="checkbox"/>

The Level 5 Data Engineer EPA has two assessment methods, each with its own mapping of KSBs. The Assessment plan and assessment guidance documents above list the criteria and KSBs that are assessed. The criteria group the KSBs and describe what the apprentice needs to do to achieve a pass or distinction for that assessment method.

Both assessment methods need to be passed by the candidate:

AM1: Project with report

AM2: Professional discussion

The End Point Assessment Organisation (EPAO) is BCS, the Chartered Institute for IT.

Assessment Part 1: Project

Learners will complete a project and write a report of 3500 words. Project brief must be submitted at the end of the practical learning period as the learner is entering the so-called “gateway” to EPA. The learner’s manager will need to approve the project brief.

- Learners will have 10 weeks to complete the project and submit the report to the EPAO.
- Learners also need to prepare and give a presentation to an independent assessor on their project.
- The presentation with questions will last at least 50 minutes. The independent assessor will ask at least 6 questions about the project and presentation.
- The project must have real business application and benefit. Candidates are expected to showcase the use of appropriate standards for sustainability, privacy and security, thoroughly document their data pipeline designs, explain the choice of relevant tooling and demonstrate operational awareness of deployment, access control, risks, and how other stakeholders may be impacted positively and negatively.

The bulk of the project is completed before gateway and is not graded on its own merit. The project evaluation is assessed, however, and must be completed after gateway. **The project output must be in the form of a report and presentation.**

The apprentice must complete a project based on any of the following:

- **a specific problem**
- **a recurring issue**
- **an idea or opportunity**

To ensure the project allows the apprentice to meet the KSBs mapped to this assessment method to the highest available grade, the EPAO must sign-off the project’s title and scope at the gateway to confirm it is suitable. The EPAO must refer to the grading descriptors to ensure that projects are pitched appropriately. **The apprentice must start the project before gateway. The project evaluation report must be completed after gateway.** The employer should ensure the apprentice has the time and resources, to plan and complete their project. The apprentice may work as part of a team to complete the project, which could include internal colleagues or technical experts. The apprentice must however, complete their project report and presentation unaided and they must be reflective of their own role and contribution. The apprentice and their employer must confirm this when the report and any presentation materials are submitted.

The report must include at least:

- an executive summary (or abstract)
- an introduction
- the scope of the project (including key performance indicators, aims and objectives)
- a project plan
- research outcomes
- data product outcomes
- project outcomes
- discussion of findings
- recommendations and conclusions
- references
- appendix containing mapping of KSBs to the report.

The independent assessor must have at least 2 weeks to review the project report and any presentation materials, to allow them to prepare questions. The apprentice must be given at least 2 weeks' notice of the presentation with questions. The presentation should cover:

- an overview of the project
- the project scope (including key performance indicators)
- summary of actions undertaken by the apprentice
- project outcomes and how these were achieved

The presentation with questions must last 50 minutes. This will typically include a presentation of 20 minutes and questioning lasting 30 minutes. The independent assessor must use the full time available for questioning. **During the presentation, the apprentice must have access to audio-visual presentation equipment, flip chart and writing and drawing materials, and a computer.**

Assessment Part 2 – Professional Discussion

Learners will have a professional discussion with an independent assessor. The professional discussion is a structured, in-depth dialogue between the apprentice and an independent assessor, focusing on the knowledge, skills, and behaviours (KSBs) relevant to the assessment method. It offers the apprentice an opportunity to provide evidence of how they've met the KSBs. It will last 80 minutes. The EPAO must give an apprentice 2 weeks' notice of the professional discussion.

We expect the candidates to demonstrate examples of working with data teams on data projects and data products, showcase ideas for future-proofing data, be clear on applying problem-solving skills, show regulatory awareness, and sensitivity towards data quality, data governance and areas for continuous improvement, both personal and organisational.

Overview of the professional discussion

The EPAO will ask ten questions, two for each of the themes:

- Data quality and performance
- Problem solving
- Regulatory Compliance
- Continuous Improvement
- Continuous professional development

A structured question bank will be developed by the assessing organisation, ensuring that questions are asked at random. We suggest you have suitable evidence that you may use to help answer the questions.

Formative submissions

The aim of the formative submissions, via TurnItIn on the Hub, of your work samples (spread throughout the programme) is for you to gather enough feedback to help you master evidencing the learning outcomes for your programme. You may then find it useful to include some of these pieces in your portfolio (helping with Assessment Method 2) that you discuss with your coach, whilst other pieces can be used to help you with your project (Assessment Method 1).

Tutors (not coaches) will be providing feedback on your work during 5 formative submissions. There will be no formal marks or grades. You can then clarify the feedback with your **coach**.

Formative submissions of your work on the programme

1. Formative Piece 1 containing two items of evidence, such as screenshots of your work
Demonstrating data quality and performance in action – 500 words
end of Month 2
2. Formative Piece 2 containing two items of evidence, such as screenshots of your work
Demonstrating problem solving and a reflection on continuous improvement - 500 words
end of Month 5
3. Formative Piece 3 containing one large piece of evidence, such as an architecture diagram
Mini-Project Submission – 750 words and a diagram
end of month 9
4. Formative Piece 4 containing evidence on how you further developed your previous submission
Mini-Presentation Submission – 7 slides with speaker notes
Mid-Month 12
5. Formative Piece 5 containing two items of evidence, such as screenshots of your work
Regulatory Compliance and CPD reflection – 500 words
Mid-Month 15

Please note that Formative Pieces 1, 2 and 5 map to Assessment Method 2 KSBs (see mappings in the next section of this handbook). These three pieces can form the basis of your Portfolio (which you discuss with your coach). Pieces 3 and 4 map to Assessment Method 1 KSBs, and can help you get started of your main project.

Here are some key points for documenting your evidence:

This qualification consists of learning around Knowledge, Skills and Behaviours (The KSBs), generally in an online classroom with a Tutor, and then applying this new learning within the workplace. There are also apply tasks in your e-learning that you should be completing. The workplace application turns the learning into knowledge, skills and behaviours that you need to evidence. When they are assessed at the end of your programme, it is the KSBs that the Assessor will be interested in (the demonstration and application, not the learning process).

The role of the portfolio of evidence

You may find that after incorporating the tutor's feedback from your formatives, you will now have a high-quality piece of evidence to incorporate into your portfolio of evidence. The **portfolio** will comprise of a small number of complete and/or discrete pieces of the apprentices very best work like mini work-based projects, enabling clear demonstration of the application of learning. The portfolio should be only the best work and should tell a story of projects/work with each project having a clear start, middle and end to include problems encountered, support received etc. Not all portfolio pieces may attract your tutor's feedback, as sometimes you may need a few more pieces than there are formatives. The portfolio is evidence that the learning has been applied in a holistic and coherent way rather than evidence that learning has taken place. Some key things to keep in mind about the portfolio:

You may find that after incorporating the tutor's feedback from your formatives, you will now have a high-quality piece of evidence to incorporate into your portfolio of evidence (NB. Not all portfolio pieces may attract your tutor's feedback, as sometimes you may need a few more pieces than there are formatives)

The **portfolio** is evidence that the learning has been applied in a holistic and coherent way rather than evidence that learning has taken place. The portfolio should be only the best work and should tell a story of projects/work with each project having a clear start, middle and end to include problems encountered, support received etc. The portfolio will comprise of a small number of complete and/or discrete pieces of the apprentices very best work like mini work-based projects, enabling clear demonstration of the application of learning.

As a guide, each portfolio piece should be between 500 and 800 words with diagrams and/or screenshots. An individual portfolio piece should take you between 20-30 hours to prepare. The following applies:

- Progress is assessed with regular checkpoints on the programme with your coach.
- Portfolio submissions are separate from formative submissions. Portfolio submissions are separate from formative submissions submitted to your tutor.
- The portfolio is not assessed directly. It is not a part of your EPA.
- The portfolio helps you demonstrate the evidence during your EPA professional discussion, which is graded.
- The portfolio helps you demonstrate the evidence of attaining your KSBs.
- The portfolio demonstrates those knowledge, skills and behaviour items that all need to be showcased via work-based evidence, i.e. tasks, projects etc. Work written up in the portfolio should have a genuine benefit.
- The portfolio should be divided into examples worked on in the workplace. Projects can be written as a narrative with the KSBs highlighted as they are covered.
- The Employer Witness Testimony can be used for sensitive/confidential evidence. Speak to your coach if you need support in collecting this evidence.
- Portfolio will typically have 5-10 pieces of evidence. Think of these as items that stand alone. Each item can be written as a narrative articulating the work and how and why it was delivered.
- Use the STAR method in writing up portfolio evidence and focus on 'I did x' not 'we'
- Make it easy for the assessor to see the KSBs evidenced – reference them
- Use the STAR method (see below) in writing up portfolio evidence and focus on 'I did x' not 'we'
- **Make it easy for yourself, and for others, to see the KSBs evidenced – reference them explicitly**
- If you cannot use the tool or technique at work, articulate what it entails to show your understanding

There will be 5 submissions of the portfolio to your **coach**. By submission, we mean presenting your portfolio to the coach during your regular meetings, and or uploading it online if instructed to do so. The coach will need to be satisfied that you have fulfilled enough KSBs for your Assessment Method 2.

The STAR technique

You can use the STAR technique for all of your submissions:



The narrative should follow the STAR technique by describing WHAT they do, WHO they do it with and HOW they do it for a range of work-based projects:

S – Set the scene – WHAT was the brief - briefly describe the situation, what they did, why they did it, factors influencing their behaviour and what the result was

T – Tasks – HOW was it to be carried out – apprentices should not try to recount everything that was said or done. They should focus on what is relevant to demonstrate competency to meet the criteria and grade requirements.

A – Activities – HOW were the tasks completed, WHAT was done, WITH WHOM was it done – Apprentices should explain the reasons for their actions and decisions, it is not just what they did but also what their thought process was

R – Result – WHAT did they learn and improve, HOW did they deliver and improve themselves, WITH WHOM did they communicate and work and fundamentally WHY did they do these activities. They are able to show how they applied knowledge and understanding by linking what they did to regulations, policies and procedures etc.

Real world evidence

Your submissions must evidence work that has been undertaken in the workplace and that has had a real business impact. Projects that use fake data or dummy projects cannot be used as evidence.

NB Simulation is not acceptable for the purpose of generating evidence. A check can be 'does the activity contribute towards the business objectives?'

Evidence - good practice and advice

"Don't tell me, show me." – Provide proof at all times with valid evidence (see Tips below)

Tips:

- **Flow charts, screen-grabs of communication, emails or records are good sources of evidence. Copies of official policies and processes, and showing how you follow them can be useful** (but you are advised to seek appropriate permissions to be able to reuse official documents).
- Use the KSB Criteria in the Assessment Plan to help you use the right language and terminology.
- Please note for the Knowledge (K) criteria we are looking for application of the knowledge in the workplace. Not just knowledge of what a term means.
- For Skills (S) criteria we are looking for how you have applied the criteria in your workplace, what you did and how you did it, and why. Be specific and write in the first person ('I did x').
- For Behaviour (B) criteria be clear about what you have done and provide evidence. Evidence for B criteria could comprise messages (seek permission, be mindful of confidentiality and anonymise personally identifiable elements), or screenshots of feedback.
- Do not share confidential or sensitive information. Please have line managers approve any screen-grabs used.

Getting to EPA - recap

- Gateway checklist reviewed (by BPP and employer)
- Gateway evidence sent to EPAO (by BPP)
- EPA conducted as per the required assessment methods (by EPAO)
- Grade and feedback provided (by EPAO)

The End Point Assessment Organisation is BCS, the Chartered Institute for IT.

Unfunded learners

- Unfunded learners will not go through external EPA.
- They will undergo an internal EPA process consisting of a professional discussion.
- They will receive a generic certificate of completion from BPP.

KSBs by Assessment Methods (Mapping Table)

Knowledge	Assessment methods
K1 Processes to monitor and optimise the performance of the availability, management and performance of data product.	Professional discussion
K2 Methodologies for moving data from one system to another for storage and further handling.	Project evaluation report, presentation and questions
K3 Data normalisation principles and the advantages they achieve in databases for data protection, redundancy, and inconsistent dependency.	Professional discussion
K4 Frameworks for data quality, covering dimensions such as accuracy, completeness, consistency, timeliness, and accessibility.	Project evaluation report, presentation and questions
K5 The inherent risks of data such as incomplete data, ethical data sources and how to ensure data quality.	Professional discussion
K6 Software development principles for data products, including debugging, version control, and testing.	Project evaluation report, presentation and questions
K7 Principles of sustainable data products and organisational responsibilities for environmental social governance.	Project evaluation report, presentation and questions
K8 Deployment approaches for new data pipelines and automated processes.	Project evaluation report, presentation and questions
K9 How to build a data product that complies with regulatory requirements.	Project evaluation report, presentation and questions
K10 Concepts of data governance, including regulatory requirements, data privacy, security, and quality control. Legislation and its application to the safe use of data.	Professional discussion
K11	Professional discussion

Knowledge	Assessment methods
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.	Project evaluation report, presentation and questions
K13 The implications of financial, strategic and compliance regarding to security, scalability, compliance and cost of local, remote or distributed solutions.	Project evaluation report, presentation and questions
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.	Project evaluation report, presentation and questions
K15 Data warehousing principles, including techniques such as star schemas, data lakes, and data marts.	Project evaluation report, presentation and questions
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.	Professional discussion
K17 Approaches to data integration and how combining disparate data sources delivers value to an organisation.	Project evaluation report, presentation and questions
K18 How to use streaming, batching and on-demand services to move data from one location to another.	Professional discussion
K19 Differences between structured, semi-structured, and unstructured data.	Project evaluation report, presentation and questions
K20 Types and uses of data engineering tools and applications in own organisation.	Project evaluation report, presentation and questions
K21 Policies and strategies to ensure business continuity for operations, particularly in relation to data provision.	Professional discussion

Knowledge	Assessment methods
K22 Technology and service management best practice including configuration, change and incident management.	Professional discussion
K23 How to undertake analysis and root cause investigation.	Professional discussion
K24 Processes for evaluating prototypes and taking them to implementation within a production environment.	Project evaluation report, presentation and questions
K25 The lifecycle of implementing data solutions in a business, from scoping, through prototyping, development, production, and continuous improvement.	Project evaluation report, presentation and questions
K26 Data development frameworks and approved organisational architectures.	Project evaluation report, presentation and questions
K27 The principles of descriptive, predictive and prescriptive analytics.	Professional discussion
K28 Continuous improvement including how to: capture good practice and lessons learned.	Professional discussion
K29 Strategies for keeping up to date with new ways of working and technological developments in data science, data engineering and AI.	Professional discussion
K30 The methods and techniques used to communicate messages to meet the needs of the audience.	Project evaluation report, presentation and questions
Skill	Assessment methods
S1 Collate, evaluate and refine user requirements to design the data product.	Project evaluation report, presentation and questions
S2 Collate, evaluate and refine business requirements including cost, resourcing, and accessibility to design the data product.	Project evaluation report, presentation and questions
S3 Design a data product to serve multiple needs and with scalability, efficiency, and security in mind.	Project evaluation report, presentation and questions

Knowledge	Assessment methods
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.	Project evaluation report, presentation and questions
S5 Produce and maintain technical documentation explaining the data product, that meets organisational, technical and non-technical user requirements, retaining critical information.	Project evaluation report, presentation and questions
S6 Systematically clean, validate, and describe data at all stages of extract, transform, load (ETL).	Project evaluation report, presentation and questions
S7 Work with different types of data stores, such as SQL, NoSQL, and distributed file system.	Professional discussion
S8 Identify and troubleshoot issues with data processing pipelines.	Professional discussion
S9 Query and manipulate data using tools and programming such as SQL and Python. Manage database access, and implement automated validation checks.	Project evaluation report, presentation and questions
S10 Communicate downtime and issues with database access to stakeholders to mitigate the operational impact of unforeseen issues.	Professional discussion
S11 Evaluate opportunities to extract value from existing data products through further development, considering costs, environmental impact and potential operational benefits.	Professional discussion
S12 Maintain a working knowledge of data use cases within organisations.	Professional discussion
S13 Use data systems securely to meet requirements and in line with organisational procedures and legislation.	Professional discussion
S14 Identify new tools and technologies and recommend potential opportunities for use in own department or organisation.	Professional discussion

Knowledge	Assessment methods
S15 Optimise data ingestion processes by making use of appropriate data ingestion frameworks such as batch, streaming and on-demand.	Professional discussion
S16 Develop algorithms and processes to extract structured data from unstructured sources.	Project evaluation report, presentation and questions
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.	Professional discussion
S18 Develop simple forecasts and monitoring tools to anticipate or respond immediately to outages and incidents.	Professional discussion
S19 Identify and escalate risks with suggested mitigation/resolutions as appropriate.	Professional discussion
S20 Investigate and respond to incidents, identifying the root cause and resolution with internal and external stakeholders.	Professional discussion
S21 Identify and remediate technical debt, assess for updates and obsolescence as part of continuous improvement.	Professional discussion
S22 Develop, maintain collaborative relationships using adaptive business methodology with stakeholders such as, business users, data scientists, data analysts and business intelligence teams.	Project evaluation report, presentation and questions
S23 Present, communicate, and disseminate messages about the data product, tailoring the message and medium to the needs of the audience.	Project evaluation report, presentation and questions
S24 Evaluate the strengths and weaknesses of prototype data products and how these integrate within an organisation's overarching data infrastructure.	Project evaluation report, presentation and questions
S25 Assess and identify gaps in existing tools and technologies in respect of implementing changes required.	Professional discussion

Knowledge	Assessment methods
S26 Identify data quality metrics and track them to ensure the quality, accuracy and reliability of the data product.	Project evaluation report, presentation and questions
S27 Selects and apply sustainable solutions to contribute to net zero and environmental strategies across the various stages of product and service delivery.	Project evaluation report, presentation and questions
S28 Horizon scanning to identify new technologies that offer increased performance of data products.	Professional discussion
S29 Implement personal strategies to keep up to date with new technology and ways of working.	Professional discussion
Behaviour	Assessment methods
B1 Acts proactively and takes accountability adapting positively to changing work priorities, ensuring deadlines are met.	Project evaluation report, presentation and questions
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.	Project evaluation report, presentation and questions
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.	Professional discussion
B4 Takes personal responsibility towards net zero and prioritises environmental sustainability outcomes in how they carry out the duties of their role.	Professional discussion
B5 Use initiative and innovation to problem solve and trouble shoot, providing creative solutions.	Professional discussion
B6 Keeps abreast of developments in emerging, contemporary and advanced technologies to optimise sustainable data products and services.	Professional discussion

Project KSBs by theme

KSBS GROUPED BY THEME	Knowledge	Skills	Behaviour
<p>Data product design</p> <p>K6 K7 K9 K12 K13 K14 S1 S2 S3 S4 S5 S27 B1</p>	<p>Software development principles for data products, including debugging, version control, and testing. (K6)</p> <p>Principles of sustainable data products and organisational responsibilities for environmental social governance. (K7)</p> <p>How to build a data product that complies with regulatory requirements. (K9)</p> <p>How to cost and build a system whilst ensuring that organisational strategies for sustainable, net zero technologies are considered. (K12)</p> <p>The implications of financial, strategic and compliance regarding to security, scalability, compliance and cost of local, remote or distributed solutions. (K13)</p> <p>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. (K14)</p>	<p>Collate, evaluate and refine user requirements to design the data product. (S1)</p> <p>Collate, evaluate and refine business requirements including cost, resourcing, and accessibility to design the data product. (S2)</p> <p>Design a data product to serve multiple needs and with scalability, efficiency, and security in mind. (S3)</p> <p>Automate data pipelines such as batch, real-time, on demand and other processes using programming languages and data integration platforms with graphical user interfaces. (S4)</p> <p>Produce and maintain technical documentation explaining the data product, that meets organisational, technical and non-technical user requirements, retaining critical information. (S5)</p> <p>Selects and apply sustainable solutions to contribute to net zero and environmental strategies across the various stages of product and service delivery. (S27)</p>	<p>Acts proactively and takes accountability adapting positively to changing work priorities, ensuring deadlines are met. (B1)</p>
<p>Data product deployment and evaluation</p> <p>K2 K4 K8 K15 K17 K19 K20 K24 K25 K26 S6 S9 S16 S24 S26</p>	<p>Methodologies for moving data from one system to another for storage and further handling. (K2)</p> <p>Frameworks for data quality, covering dimensions such as accuracy, completeness, consistency, timeliness, and accessibility. (K4)</p> <p>Deployment approaches for new data pipelines and automated processes. (K8)</p>	<p>Systematically clean, validate, and describe data at all stages of extract, transform, load (ETL). (S6)</p> <p>Query and manipulate data using tools and programming such as SQL and Python. Manage database access, and implement automated validation checks. (S9)</p> <p>Develop algorithms and processes to extract structured</p>	<p>None</p>

KSBS GROUPED BY THEME	Knowledge	Skills	Behaviour
	<p>Data warehousing principles, including techniques such as star schemas, data lakes, and data marts. (K15)</p> <p>Approaches to data integration and how combining disparate data sources delivers value to an organisation. (K17)</p> <p>Differences between structured, semi-structured, and unstructured data. (K19)</p> <p>Types and uses of data engineering tools and applications in own organisation. (K20)</p> <p>Processes for evaluating prototypes and taking them to implementation within a production environment. (K24)</p> <p>The lifecycle of implementing data solutions in a business, from scoping, through prototyping, development, production, and continuous improvement. (K25)</p> <p>Data development frameworks and approved organisational architectures. (K26)</p>	<p>data from unstructured sources. (S16)</p> <p>Evaluate the strengths and weaknesses of prototype data products and how these integrate within an organisation's overarching data infrastructure. (S24)</p> <p>Identify data quality metrics and track them to ensure the quality, accuracy and reliability of the data product. (S26)</p>	
<p>Collaborative working K30 S22 S23 B2</p>	<p>The methods and techniques used to communicate messages to meet the needs of the audience. (K30)</p>	<p>Develop, maintain collaborative relationships using adaptive business methodology with stakeholders such as, business users, data scientists, data analysts and business intelligence teams. (S22)</p> <p>Present, communicate, and disseminate messages about the data product, tailoring the message and medium to the needs of the audience. (S23)</p>	<p>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. (B2)</p>

Professional discussion KSBs by theme

KSBS GROUPED BY THEME	Knowledge	Skills	Behaviour
Data quality and performance K1 K3 K5 K18 K27 S7 S15	<p>Processes to monitor and optimise the performance of the availability, management and performance of data product. (K1)</p> <p>Data normalisation principles and the advantages they achieve in databases for data protection, redundancy, and inconsistent dependency. (K3)</p> <p>The inherent risks of data such as incomplete data, ethical data sources and how to ensure data quality. (K5)</p> <p>How to use streaming, batching and on-demand services to move data from one location to another. (K18)</p> <p>The principles of descriptive, predictive and prescriptive analytics. (K27)</p>	<p>Work with different types of data stores, such as SQL, NoSQL, and distributed file system. (S7)</p> <p>Optimise data ingestion processes by making use of appropriate data ingestion frameworks such as batch, streaming and on-demand. (S15)</p>	None
Problem Solving K21 K22 K23 S8 S10 S12 S18 S19 S20 B5	<p>Policies and strategies to ensure business continuity for operations, particularly in relation to data provision. (K21)</p> <p>Technology and service management best practice including configuration, change and incident management. (K22)</p> <p>How to undertake analysis and root cause investigation. (K23)</p>	<p>Identify and troubleshoot issues with data processing pipelines. (S8)</p> <p>Communicate downtime and issues with database access to stakeholders to mitigate the operational impact of unforeseen issues. (S10)</p> <p>Maintain a working knowledge of data use cases within organisations. (S12)</p> <p>Develop simple forecasts and monitoring tools to anticipate or respond immediately to outages and incidents. (S18)</p> <p>Identify and escalate risks with suggested mitigation/resolutions as appropriate. (S19)</p> <p>Investigate and respond to incidents, identifying the root cause and resolution with internal and external stakeholders. (S20)</p>	Use initiative and innovation to problem solve and trouble shoot, providing creative solutions. (B5)

KSBS GROUPED BY THEME	Knowledge	Skills	Behaviour
Regulatory Compliance K10 K11 S13	<p>Concepts of data governance, including regulatory requirements, data privacy, security, and quality control. Legislation and its application to the safe use of data. (K10)</p> <p>Data and information security standards, ethical practices, policies and procedures relevant to data management activities such as data lineage and metadata management. (K11)</p>	Use data systems securely to meet requirements and in line with organisational procedures and legislation. (S13)	None
Continuous Improvement K16 K28 S11 S14 S17 S21 S25 S28 B3 B4	<p>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. (K16)</p> <p>Continuous improvement including how to: capture good practice and lessons learned. (K28)</p>	<p>Evaluate opportunities to extract value from existing data products through further development, considering costs, environmental impact and potential operational benefits. (S11)</p> <p>Identify new tools and technologies and recommend potential opportunities for use in own department or organisation. (S14)</p> <p>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. (S17)</p> <p>Identify and remediate technical debt, assess for updates and obsolescence as part of continuous improvement. (S21)</p> <p>Assess and identify gaps in existing tools and technologies in respect of implementing changes required. (S25)</p> <p>Horizon scanning to identify new technologies that offer increased performance of data products. (S28)</p>	<p>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. (B3)</p> <p>Takes personal responsibility towards net zero and prioritises environmental sustainability outcomes in how they carry out the duties of their role. (B4)</p>
Continuous professional development	Strategies for keeping up to date with new ways of working and technological developments	Implement personal strategies to keep up to date with new	Keeps abreast of developments in emerging, contemporary and

KSBS GROUPED BY THEME	Knowledge	Skills	Behaviour
K29 S29 B6	in data science, data engineering and AI. (K29)	technology and ways of working. (S29)	advanced technologies to optimise sustainable data products and services. (B6)

Grading basis

Project evaluation report, presentation and questions

Fail - does not meet pass criteria

Theme KSBs	Pass Apprentices must demonstrate all of the pass descriptors	Distinction Apprentices must demonstrate all of the pass descriptors and all of the distinction descriptors
Data product design K6 K7 K9 K12 K13 K14 S1 S2 S3 S4 S5 S27 B1	<p>Demonstrates how they have collated, evaluated and refined user requirements to design and build a scalable data product that serves multiple needs and complies with regulatory requirements. (K9, S1, S3)</p> <p>Explains how they collated, evaluated and refined business requirements, to design, build and maintain a system whilst ensuring that organisational strategies for sustainable, net-zero technologies are considered. (K12 & S2)</p> <p>Explains how they selected sustainable solutions in relation to data products and environmental social governance to ensure the use of less carbon across the various stages of product and service delivery. (K7, S27)</p> <p>Demonstrates how they used security, scalability and governance when automating data pipelines using programming languages and data integration platforms with graphical user interfaces. (K13, S4)</p> <p>Demonstrates how they have taken accountability produced and maintained technical documentation for a data product in order to meet organisational user requirements, whilst adapting to changing work priorities to ensure that deadlines are met. (S5, B1)</p> <p>Explains how debugging, version control and testing have an impact on software development and the principles for data products. (K6)</p> <p>Outlines the uses of different on-demand cloud computing platforms. (K14)</p>	Justifies how the data product created met the requirements and served multiple needs (S1, S3)
Data product deployment and evaluation K2 K4 K8 K15 K17 K19 K20 K24 K25 K26 S6 S9 S16 S24 S26	<p>Explains the deployment approaches processes for new data pipelines and automated processes.(K8)</p> <p>Explains techniques such as star schemas, data lakes and data marts and the impact they have on data warehousing principles. (K15)</p> <p>Demonstrate how to systematically clean, validate and describe data at all stages of extract, transform and load, showing how combining disparate data sources and taking different approaches to data integration delivers value to an organisation. (K17, S6)</p> <p>Describes the types and uses of data engineering tools in their own organisation and how they apply them. (K20)</p> <p>Evaluates the strengths and weaknesses of prototype data products to integrate within an organisation's overarching data structure, taking into consideration the lifecycle of implementing data solutions in a business. (K24, K25, S24)</p> <p>Describes the approved organisational architectures and the relevant data development frameworks. (K26)</p> <p>Identifies data quality metrics and their frameworks and tracks them to ensure quality, accuracy and reliability of the data product. (K4, S26)</p> <p>Demonstrates the use of tools and programming to query and manipulate data and implement automated validation checks, showing the methodologies used for moving data from one system to another for storage and handling. (K2, S9)</p>	Evaluates the success of the algorithm developed (S16)

Theme KSBs	Pass Apprentices must demonstrate all of the pass descriptors	Distinction Apprentices must demonstrate all of the pass descriptors and all of the distinction descriptors
	Explains how they have worked with structured, semi-structured and unstructured data, developing algorithms to extract from sources (K19, S16)	
Collaborative working K30 S2 2 S23 B2	<p>Outlines the methods and techniques used to communicate messages about the data product that meet the needs of the audience. (K30, S23)</p> <p>Explains how they worked collaboratively with different technical and non-technical stakeholders, using adaptive business methodology to support an inclusive culture and develop and maintain strong working relationships in order to achieve common goals. (S22, B2)</p>	Evaluate the impact of the methods and techniques used to communicate messages about the data product to the audience. (K30, S23)

Professional discussion

Fail - does not meet pass criteria

Theme KSBs	Pass Apprentices must demonstrate all of the pass descriptors	Distinction Apprentices must demonstrate all of the pass descriptors and all of the distinction descriptors
Data quality and performance K1 K3 K5 K18 K27 S7 S15	<p>Explains how they monitor different types of data store to optimise system management, performance and availability. (K1, S7)</p> <p>Defines data normalisation principles and the advantages that they achieve for data protection, redundancy and inconsistent dependency. (K3)</p> <p>Explains the inherent risks of data and how to ensure data quality (K5)</p> <p>Explains the principles of descriptive, predictive and prescriptive analytics. (K27)</p> <p>Describes how they use data ingestion frameworks such as streaming, batching and on demand services to move data from one location to another in order to optimise data ingestion processes. (K18, S15)</p>	Compares and contrasts the different types of data stores they have used and how they optimised performance (K1, S7)
Problem Solving K21 K22 K23 S8 S10 S12 S18 S19 S20 B5	<p>Describes technology and service management best practice. (K22)</p> <p>Explains how they identify and escalate risks and incidents, communicating downtime and issues with database access in line with policies in order to mitigate operational impact whilst ensuring business continuity. (K21, S10, S18, S19)</p>	Justifies the approach taken to manage risks and incidents to maintain business continuity. (S18, S19)

Theme KSBs	Pass Apprentices must demonstrate all of the pass descriptors	Distinction Apprentices must demonstrate all of the pass descriptors and all of the distinction descriptors
	<p>Explains how they have maintained a working knowledge of data use cases within organisations. (S12)</p> <p>Explains how their analysis of root cause investigation is used to respond to incidents within data processing pipelines, whilst troubleshooting and providing resolutions to stakeholders. (K23, S8, S20, B5)</p>	
Regulatory Compliance K10 K11 S13	<p>Explains their use of data, information security standards, ethical practices and data management policies and procedures to ensure data systems are used securely and in accordance with relevant legislation. (K11, S13)</p> <p>Explains the legislative associated with the use and collation of data, including concepts of data governance and regulatory requirements. (K10)</p>	None
Continuous Improvement K16 K28 S11 S14 S17 S21 S25 S28 B3 B4	<p>Outlines how they evaluate opportunities to extract value from existing data products whilst applying the principles of data and considering costs, environmental impact and potential operating benefits. (K16, S11)</p> <p>Explains how they take personal responsibility within the duties of their role to identify new tools and technologies, and recommend potential opportunities for use in own department or organisation in order to prioritise environmental sustainability outcomes to work towards net zero. (S14, B4)</p> <p>Explains how they take a quality focussed approach to identify and remediate technical debt and assess for updates and obsolescence within their promotion of continuous improvement, by utilising peer review techniques and capturing good practice, to provide innovation and creativity to the data system development process in order to improve processes and address business challenges. (K28, S21, B3)</p> <p>Explains how they apply ways of working that support software development principles and advocate software development best practice when working with other data professionals. (S17)</p> <p>Explains how they identify and assess new technologies, as well as gaps in existing tools and technologies, that offer increased performance of data products and implementation of changes required. (S25, S28)</p>	Evaluates the impact that the implementation of identified new technologies would have on practices within the organisation. (S25, S28)
Continuous professional development K29 S29 B6	Explains how they have implemented personal strategies for keeping up to date with new ways of working and to keep abreast of developments in emerging, contemporary and advanced technologies, in order to keep up to date with new technologies and technological developments in data science, data engineering and AI and to optimise sustainable products and services (K29, S29, B6)	Evaluate the impact that keeping up to date with technological developments has had on their own professi

Grades available

The grades available for each assessment method are below:

Project evaluation report, presentation and questions:

- fail
- pass
- distinction

Professional discussion:

- fail
- pass
- distinction

Overall EPA and apprenticeship can be graded:

- fail
- pass
- merit
- distinction

Overall EPA grading

All EPA methods must be passed for the EPA to be passed overall. Apprentices must gain a pass in one method plus a pass in the other method to gain a pass overall. Apprentices must gain a pass in one method plus a distinction in the other method to gain a merit overall. Apprentices must gain a distinction in both assessment methods to gain a distinction overall. Grades from individual assessment methods should be combined in the following way to determine the grade of the EPA as a whole:

Assessment Method 1	Assessment Method 2	Overall Grading
Project	Professional discussion	
Fail	Any grade	Fail
Any grade	Fail	Fail
Pass	Pass	Pass
Pass	Distinction	Merit
Distinction	Pass	Merit
Distinction	Distinction	Distinction

Re-sits and re-takes

Re-sit and re-take policy is outlined in the IfATE Assessment Plan as follows:

Apprentices who fail one or more assessment method will be offered the opportunity to take a re-sit or a re-take at the employer's discretion. The apprentice's employer will need to agree that either a re-sit or re-take is an appropriate course of action.

A re-sit does not require further learning, whereas a re-take does.

Apprentices should have a supportive action plan to prepare for a re-sit or a re-take.

An apprentice who fails one or more assessment methods, and therefore the EPA in the first instance, will be required to re-sit or re-take the failed assessment method(s) only.

The timescales for a re-sit/re-take is agreed between the employer and EPAO. A re-sit is typically taken within two months of the EPA outcome notification. The timescale for a re-take is dependent on how much re-training is required and is typically taken within four months of the EPA outcome notification.

All assessment methods must be taken within a six-month period, otherwise the entire EPA will need to be re-sat/re-taken.

Re-sits and re-takes are not offered to apprentices wishing to move from pass to a higher grade.

The overall grade awarded for this apprenticeship standard is not capped, meaning that an apprentice can still achieve a maximum EPA grade of distinction if any assessment method needs to be re-sat or re-taken.

Programme roles

Tutors deliver the taught content on the programme, via live webinars online. You can communicate with tutors via the forum where they can post and open discussions with their cohort.

Coaches work one to one with you to guide you through the programme, monitoring the time off the job, and guiding you through their portfolio development, assignments and into Gateway and EPA.

Module leads are responsible for the content and materials of individual modules. Module leads manage content on the Hub.

The **Award Lead (Programme Lead)** is responsible for the content and curriculum of the programme.

The **Operation Coordinator team** manage learner communication, enrolments, room access and learner schedules.

New Funding Rules 22/23

As of the 1st August 2022 the ESFA created new rules that govern apprenticeship programmes, including the Active Learning Rule.

Active Learning Rule

The Active Learning Rule states that all learners, whilst on programme, must be involved in active learning throughout their practical learning period. It means that there must be evidence of learning every 4 weeks whilst the apprentice is on programme during the practical learning period.

What constitutes evidence?

The main types of learning that could be considered as evidence are:

- Attending webinars/lectures
- Completing a learning activity on the HUB
- Submission of an assignment
- Off the job training
- Self-Declaration

Off the job training rule

Off the Job (OTJ) Training is a rule set by the Education and Skills Funding Agency (ESFA) which states that the apprentice must spend time, whilst on programme, learning and developing new knowledge, skills and behaviours.

What is OTJT

It is training which is received by the apprentice within their **practical period**, during the apprentice's **normal working hours**, for the purpose of **achieving the knowledge, skills and behaviours** of the apprenticeship. By normal working hours we mean the hours for which the apprentice would normally be paid, excluding overtime.

New rules August 2022 onwards:

From DfE Apprenticeship Funding Rules 2022-2023

"We have amended the off-the-job training policy to remove the link to the apprentice's working hours (for full-time apprentices). The new minimum off the-job training requirement for a full-time apprentice is 20% of a 30 hour week (even where the apprentice works more than 30 hours per week for an employer); this equates to an average of 6 hours of off-the-job training per week. The minimum requirement for a 19 part-time apprentice remains unchanged (i.e. 20% of their normal working hours over an extended duration). This policy change only applies to new starts from 1 August 2022." (P43):

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1109627/2223_Provider_Rules_Version_2_Final.pdf

For learners prior to this start date:

- A minimum of 20% of a learner's working hours throughout their Apprenticeship should be spent on training and development, relevant to the achievement of their programme.

Off the job time is tracked in a tracker provided to the learner by the coach.

Additional Guidance From BCS (the EPAO) is Available Online

There is a range of resources available for apprentices to explore what's involved in each assessment method and how you can best prepare for success. These are available in the apprenticeship section within MyBCS. Apprentices who have not been registered for BCS membership, can create an account for free, which will then give them access to MyBCS.

List of all Knowledge Skills & Behaviours (KSBs)

KSBs, short for "Knowledge, Skills, and Behaviours," are the fundamental attributes apprentices need to develop to excel in their chosen occupation. These attributes complement the technical studies and exams apprentices undertake and form the primary basis for their end-point assessment (EPA). They can be thought of as interpersonal and practical skills that are vital in a workplace setting. In essence, KSBs serve as the cornerstone of an apprentice's comprehensive development, encompassing technical knowledge, practical skills, and professional conduct that collectively contribute to a successful and well-rounded career.

We have made sure that the KSBs for the programme are mapped onto the relevant learning outcomes to make them less overwhelming for the learner. It is still advisable that the learner looks through the list of KSBs for this programme (they do not need to be memorised!).

Knowledge

The knowledge aspect of KSBs pertains to the educational aspects of the apprenticeship journey. It encompasses the information and technical understanding necessary to perform the role effectively. This knowledge can be specific to the pursued occupation and encompass more general concepts. In certain cases, this knowledge might align with technical qualifications in fields such as accountancy or data analysis.

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, through 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

These encompass the practical application of acquired knowledge to the job at hand. These skills are cultivated through a combination of hands-on and theoretical training. The skills must learn to go beyond routine tasks, instead focusing on bridging gaps in expertise and enhancing proficiency.

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 either 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

The behaviours component centres on an apprentice's mindset and work ethic. It involves adopting a manner of conduct that aligns with the standards and expectations of the organisation and the profession. While some behaviours might come naturally, many can be learned and improved. Behaviours are often quite adaptable and transferrable between different roles. Examples include teamwork, adaptability, and maintaining a high level of professionalism.

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 none 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.

Maths and English on the Programme

Recognising the fundamental role that proficient communication and numerical skills play in professional and personal success, our curriculum is designed to develop these essential competencies through embedded learning opportunities.

Functional Skills Icons

Throughout this programme, specific sections of the curriculum that focus on English and maths skills are marked with the "Focus on Functional Skills" icon. This visual marker is an easy reference for learners to identify key learning segments where functional skills are taught explicitly or applied within the course content.

English Skills Development

The ability to communicate effectively is crucial in the field of data engineering. As such, our curriculum includes targeted learning outcomes that enhance subject-specific vocabulary and comprehension. For instance:

During webinars on data governance, learners will engage in exercises that improve technical language acquisition, critical for writing clear documentation and effective communication with stakeholders.

Interactive online content includes sessions on crafting precise data definitions and annotations, which reinforce clarity and accuracy in written communication.

Maths Skills Development

Mathematical competencies are equally vital, as they enable our learners to perform accurate analysis and resource planning. Examples of maths skills development in our curriculum include:

In sessions on database design and data storage solutions, learners will perform calculations related to data throughput and storage needs, helping them understand and plan for infrastructure requirements.

Our modules on data pipeline construction will include practical exercises in algorithm development, where learners apply mathematical logic to optimise data flow and processing.

Importance of Functional Skills

Functional skills in English and maths are not merely academic requirements; they are critical to achieving success in higher education and future careers. These skills facilitate:

Higher Education Readiness: Proficiency in English and maths ensures that learners can cope with complex texts and advanced research, as well as manage the quantitative aspects of higher academic work.

Career Competence: In the workplace, strong communication and analytical skills enable professionals to manage projects efficiently, maintain precision in technical tasks, and collaborate effectively with diverse teams.

Personal Success: Beyond professional life, these skills enhance an individual's ability to manage daily activities that involve financial decisions, problem-solving, and effective communication.

The deliberate integration of English and maths within our Data Engineering curriculum ensures that our learners are not only technically skilled but also well-equipped with the functional competencies required for comprehensive professional and personal success. The "Focus on Functional Skills" icon in our programme materials serves as a constant reminder of our commitment to developing these essential skills, ensuring that our learners are well-prepared to meet the challenges of their academic and career aspirations.

Safeguarding, Prevent and British Values

The following elements of the programme have been implemented to foster a secure, respectful, and inclusive educational setting.

Safeguarding

We are committed to safeguarding and ensuring the welfare of all our learners. To this end, our School enacts:

Policies and Procedures: Comprehensive safeguarding policies that are readily accessible, detailing the steps to take in case of concerns about a learner's welfare.

Support Networks: Dedicated welfare officers are available to provide support and guidance, ensuring that all learners feel safe and supported throughout their educational journey.

If you are concerned or would like further information about Safeguarding within BPP please contact

safeguarding@bpp.com

or visit <https://www.bppstudents.com/support/safeguarding/>

All BPP support services can be found at <https://www.bppstudents.com/support/>

Training and Resources: Regular training sessions for both staff and learners on safeguarding principles, aimed at recognising, responding to, and reporting potential welfare issues.

Prevent Strategy

In adherence to the Prevent duty, our curriculum is designed to:

Promote Critical Thinking: Encourage a questioning attitude towards information and foster critical analysis skills that are essential in countering extremist narratives.

Monitoring and Intervention: Implement procedures to identify at-risk individuals and intervene appropriately in a supportive and respectful manner.

British Values

The promotion of British Values is integral to our curriculum, encompassing:

Democracy: Encouraging learners to understand and engage with democratic processes and how they can influence decision-making, both within and outside the educational environment.

Rule of Law: Highlighting the importance of laws in governing behaviour, ensuring that learners understand the reasons behind laws and the consequences of breaking them.

Individual Liberty and Mutual Respect: Supporting learners to develop their self-knowledge, self-esteem, and self-confidence. Fostering an environment of mutual respect and tolerance for those with different faiths and beliefs.

Tolerance: Activities and discussions designed to provide learners with the opportunity to encounter cultural diversity and to understand the value of tolerance in a pluralistic society.

The integration of safeguarding, the Prevent strategy, and British Values into our Data Engineering programme reflects our commitment to not only educate but also to cultivate a responsible, respectful, and socially aware cohort of professionals. Our programme aims to equip learners with the knowledge and values needed to thrive in a diverse and challenging world, upholding the highest standards of safety, integrity, and respect.

Core Reading List

The learning on the BPP L5 Data Engineering course will be strengthened by the learner following these reading materials written by authors with experience of the industry.

- Baum, D. (2020). *Cloud Data Engineering For Dummies®*, Snowflake Special Edition, Wiley and Sons [10 hours]
- Beattie, T. (2020). *DevOps Culture and Practice with OpenShift*. Packt Publishing [30 hours]
- Been, H. (2020). *Implementing Azure DevOps Solutions*. Packt Publishing [30 hours]
- Campbell, L., Majors, C. (2017). *Database Reliability Engineering*, O'Reilly Media [20 hours]
- Chou, E. (2020). *Mastering Python Networking*. Packt Publishing [20 hours]
- Databricks (2023), *The Big Book of Data Engineering* 2nd Edition, Databricks Online eBook (free) [15 hours]
- Dehghani, Z. (2022) *Data Mesh: Delivering Data-Driven Value at Scale*. O'Reilly Media. [25 hours]
- Densmore, J. (2021). *Data pipelines pocket reference*. O'Reilly Media. [15 hours]
- Done, P., Kamsky, A. (2023). *Practical MongoDB Aggregations*. Packt Publishing [25 hours]
- Eagar, G. (2021). *Data Engineering with AWS*. Packt Publishing [40 hours]
- Eryurek, E., Gilad, U. (2021). *Data Governance: The Definitive Guide*. O'Reilly [25 hours]
- Espenschutz, G. (2023). *Data Ingestion with Python – Cookbook*. Packt Publishing [30 hours]
- Grigorev, A. (2024). <https://github.com/DataTalksClub/data-engineering-zoomcamp> [40 hours]
- Harenslak, B, De Ruiter, JR, (2021) *Data Pipelines with Apache Airflow*. Manning [30 hours]
- Haskell, C. (2024). *Driving Data Projects: A comprehensive guide*. British Computer Society [20 hours]
- Haviv, Y. (2024). *Implementing MLOps in the Enterprise*. O'Reilly. [20 hours]
- Hyman, J.A. (2023). *Microsoft Azure for Dummies*. [20 hours]
- Jhajj R (2023). *Mastering MLOps Architecture*. BPB Publications [15 hours]
- Kaufmann, M. (2022). *Accelerate DevOps with GitHub*. Packt Publishing. [30 hours]
- King, T., Schwarzenbach, J., (2020). *Managing Data Quality: A practical guide*. British Computer Society [15 hours]
- Labouardy, M. (2021). *Pipeline as Code. Continuous Delivery with Jenkins, Kubernetes and Terraform*. Manning. [40 hours]
- Leszko, R. (2022). *Continuous Delivery with Docker and Jenkins*. Packt Publishing. [30 hours]
- Luksa, M. (2018) *Kubernetes in Action*. Manning [20 hours]
- McMahon, A. (2023). *Machine Learning Engineering with Python*. Packt Publishing [40 hours]
- Mertz, D. (2021) *Cleaning Data for Effective Data Science*. Packt Publishing. [20 hours]
- Perkins, L., Redmond, E., Wilson, J. (2018) *Seven Databases in Seven Weeks*. O'Reilly Media [25 hours]
- Pote, S. (2023). *Machine Learning in Production*. BPB Publications. [30 hours]
- Reis, J., Housley, M. (2022). *Fundamentals of Data Engineering: Plan and Build Robust Data Systems*. O'Reilly [25 hours]
- Shrivastava, S. (2023). *AWS for Solutions Architects*. Packt Publishing. [35 hours]
- Simsion, G., Witt, G. (2005) *Data Modeling Essentials* 3rd ed., Elsevier [20 hours]
- Skoulkari, A. (2023) *Learning Git*. O'Reilly. [20 hours]
- Taze, M. (2021) *Engineers Survival Guide*. Self-published. Available on Amazon. [10 hours]
- Zhao, A. (2021). *SQL Pocket Guide: A Guide to SQL Usage*. O'Reilly Media [20 hours]

Learners may log the hours spent on successful completion of their reading towards their L5 programme self-guided element by writing a book review (one-pager) in their learning journal and presenting it to their coach (max. 150 hours)

Short Industry Certifications

The content of the BPP L5 Data Engineering course will put students in a good position to attempt sitting the following self-guided short certifications provided by the industry.

- Cisco Academy NDG Linux Unhatched certificate of completion (8 hours) [free]
- Cisco Academy NDG Linux Essentials certificate of completion (70 hours) [free]
- Learning Bash Scripting certification (2 hours) [free with LinkedIn Learning]
- Microsoft Azure for Data Engineering (6 hours) [enroll for free]
- Data Storage in Microsoft Azure (16 hours) [enroll for free]
- Data Integration with Microsoft Azure Data Factory (16 hours) [enroll for free]
- Microsoft Azure Databricks for Data Engineering (21 hours) [enroll for free]
- IBM Data Engineering Foundations Specialization (90 hours) [enroll for free]
- IBM Introduction to Data Engineering (12 hours) [enroll for free]
- Mastering Data Governance and Ethics badge of completion (10 hours) [free with LinkedIn Learning]
- Cisco Introduction to Cybersecurity certification (15 hours) [free]
- Cisco Networking Essentials certification (70 hours) [free]
- Cisco Cybersecurity Essentials certification (30 hours) [free]
- Cisco PCAP Programming Essentials in Python certification (75 hours) [free]
- Getting Started with AWS storage [3 hours]
- Planning Large Scale Data Migrations to AWS [2 hours]
- Performing a Basic Audit of your AWS Environment [2 hours]
- Fundamentals of Analytics on AWS – Part 2 [3 hours]
- Digital Classroom - Cloud Operations on AWS [24 hours]

Learners may log the hours spent on successful certifications towards their L5 programme self-guided element by speaking to their coach (max. 150 hours)

L5 Data Engineer

Sign (learner)

Sign (manager)

Total self-study hours logged:

[illegible]

[illegible]