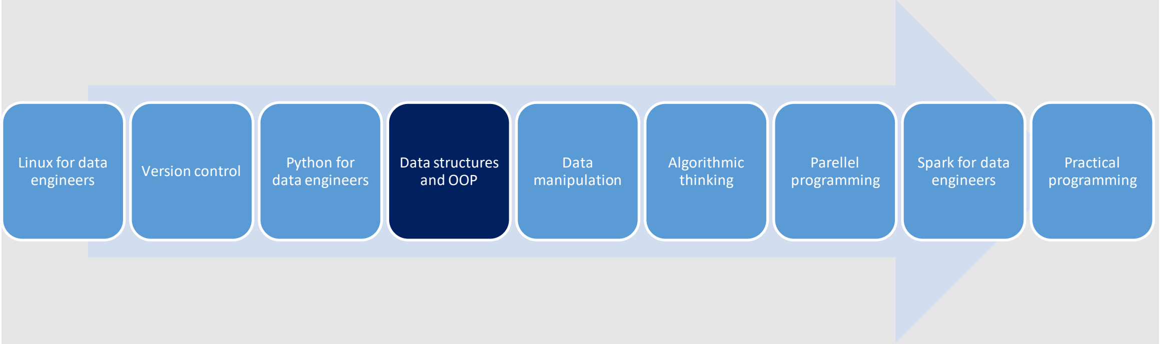
Topic 4 – Data structures and OOP

This document is the handbook for Topic 4 – **Data structures and OOP** – within Module 3 – **Programming and Scripting Essentials**.

The purpose of this document is to guide your learning throughout this topic and help you to maximise the value you get from the materials provided by the BPP School of Technology.

Context

This handbook is for one of 9 topics for this Module.



Every topic contributes towards the ultimate learning objectives for the Module, which you will be assessed on at the end of the term.

Module Learning Outcomes

On successfully completing this module, you will be able to:

- **Employ** software development tools and techniques for designing, deploying and maintaining secure data products and pipelines, including debugging, version control and testing.
- **Construct** algorithms that correctly and efficiently handle data at scale whilst mitigating risks.
- **Demonstrate** the knowledge of the steps needed to prepare the code for production.

Module Assessment

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:

(1) Project with report

The learner will complete a project and write a report of 3500 words. Project brief submitted at gateway:

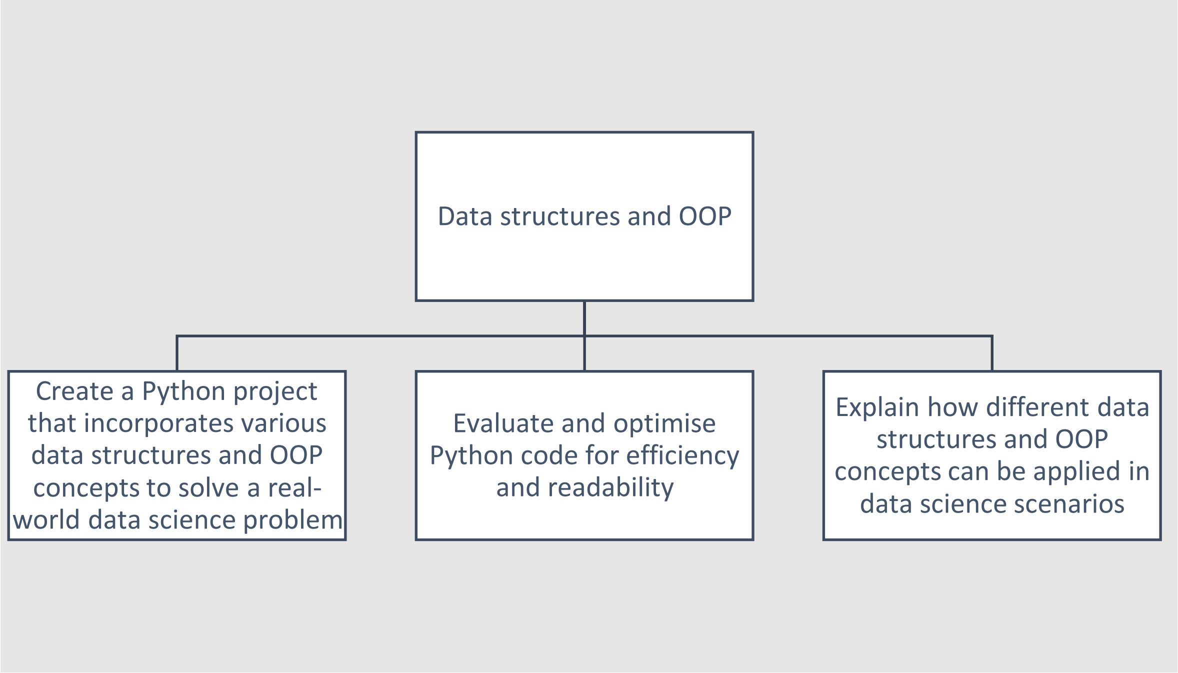
- Learners will have 10 weeks to complete the project and submit the report to the EPAO
- Learners also need to prepare and give a presentation to an independent assessor on their project
- The presentation with questions will last at least 50 minutes. The independent assessor will ask at least 6 questions about the project and presentation
- The project has to 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

(2) Professional discussion underpinned by a portfolio of evidence

- Learners will have a professional discussion with an independent assessor. It will last 80 minutes
- They will be asked at least 10 questions about Data Engineering
- The portfolio of evidence will be used to help answer the questions
- 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

Topic Learning Outcomes

As a step towards build your skills towards the final module assessment, the learning objectives for this topic are:



Data structures and OOP

Create a Python project that incorporates various data structures and OOP concepts to solve a real-world data science problem

Evaluate and optimise Python code for efficiency and readability

Explain how different data structures and OOP concepts can be applied in data science scenarios

Introduction

In today's data-driven world, organisations grapple with ever-increasing volumes of data, diverse data types, and the need for real-time processing. As a data engineer, you're tasked with creating robust data pipelines and systems to handle these demands efficiently. Without a deep understanding of data structures and OOP principles, it's easy to create solutions that are difficult to scale, maintain, or adapt to changing business needs.

Consider a global e-commerce platform processing millions of transactions daily. The company needs a system to ingest and process transaction data in real-time, provide instant product recommendations, analyse sales trends across regions and categories, and ensure data integrity and security throughout the pipeline. This scenario exemplifies why advanced knowledge of data structures and OOP is crucial for data engineers.

By the end of this topic, you will have gained a deep understanding of Python data structures, OOP concepts, and their application in data engineering. You'll comprehend best practices for code organisation and data serialisation, and understand how to design scalable, maintainable data systems. Your skills will include selecting and implementing appropriate data structures for various tasks, designing object-oriented solutions for complex data problems, writing clean and efficient Python code, and applying OOP principles to create flexible, reusable data pipeline components.

Through a blend of theoretical learning and hands-on practice, this topic will enable you to develop the expertise to tackle real-world data engineering challenges. You'll learn to create data systems that are not just functional, but also efficient, scalable, and adaptable to evolving business needs. This knowledge will equip you to design sophisticated data architectures, optimise processing workflows, and contribute more effectively to your organisation's data strategy. Whether you're working on big data processing, building ETL pipelines, or developing data APIs, the insights gained from this topic will be invaluable in your role as a data engineer.

Structure

Topics for this programme follow a Prepare-Collaborate-Apply structure:

Prepare

This is the stage where you build the knowledge to underpin your learning. This might involve completing interactive e-learning packages, watching videos, or working through reading materials.

It is essential that you make the most of the learning materials provided before attending webinars, as this will allow you to test your knowledge and stretch you understanding further.

Collaborate

This is where you will receive guidance from our expert tutors and coaches to shape and refine your understanding through in-depth explanation, discussion, testing and carrying out more advanced practical and realistic tasks. This also helps to develop valuable team-working skills.

Apply

You now apply the knowledge you have developed to real-world tasks.

Off-the-job learning tasks

This stage is all about ensuring you truly grasp and retain what you've learned. Through completion of off-the-job (OTJ) revision tasks and tests, you'll get plenty of practice applying your knowledge. Plan to dedicate 6-8 hours each week to guided study and portfolio work, with sessions typically on the same day each week.

Task 1 brief: Introduction to debugging

In this task, **you** will explore the critical skill of debugging in software development, including:

1. **Importance of Debugging**: You'll learn why debugging matters. It's about identifying and fixing errors in your code. Think of it as detective work for programmers!

2. **Python Debugging Fundamentals**:

    - **Print Statements**: You'll use print statements to track what's happening in your code. It's like leaving breadcrumbs to follow your program's journey.

    - **Python Debugger (pdb)**: This powerful tool lets you pause your program, inspect variables, and figure out what's going wrong step by step.

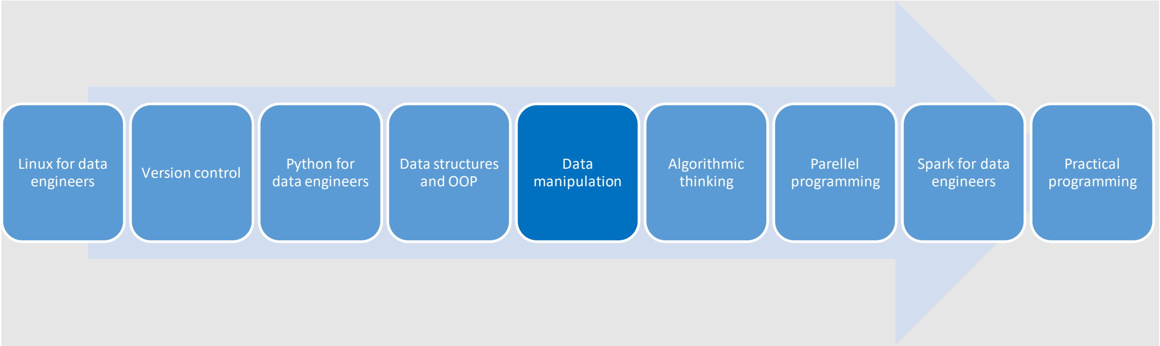
- **Exception Handling**: You'll catch and handle errors gracefully, preventing your program from crashing.
- **Logging**: Imagine keeping a journal of your program's adventures. Logging helps you record messages about what's going on.

3. **Activity**: You'll tackle a practical challenge. There's a function called stripped_reversed_lowercase with a bug. Your mission? Debug it! Use the debugger, print variables, and solve the mystery.

Further guidance can be found here: Apply e-learning: Introduction to debugging

Link

This handbook is for one of 9 topics for this Module.



The sequence of topics in this module is carefully designed so that your knowledge and skills will develop as you progress.

The next topic is **Data manipulation**.