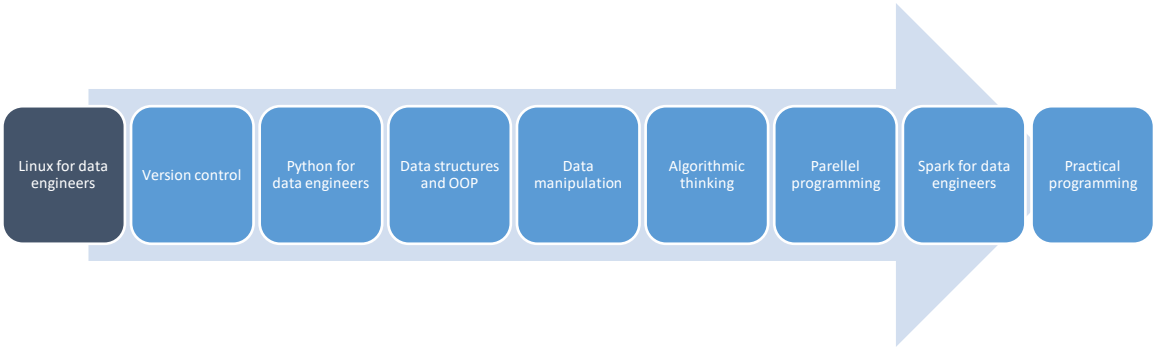
Topic 1 – Linux for data engineers

This document is the handbook for Topic 1 –**Linux for data engineers** – 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:

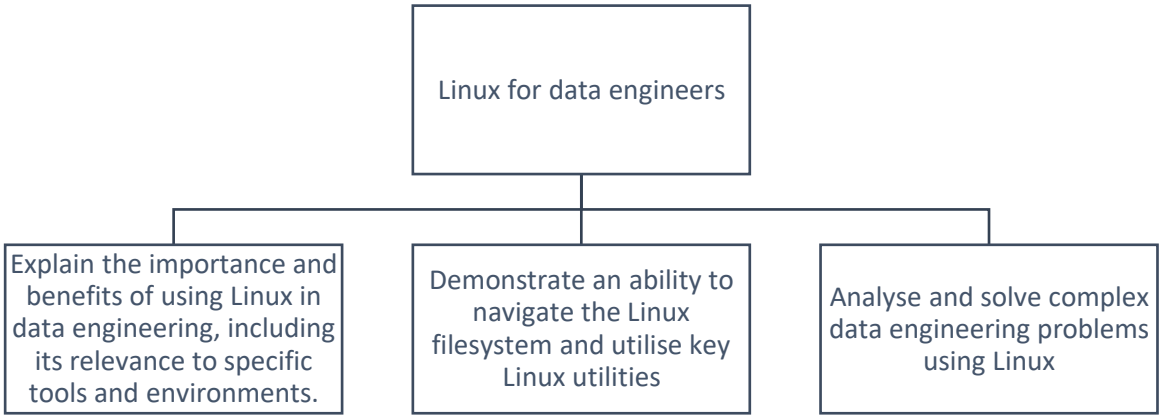
- Learners will have 10 weeks to complete the project and submit the report to the EPAO
- Learners also need to prepare and give a presentation to an independent assessor on their project
- The presentation with questions will last at least 50 minutes. The independent assessor will ask at least 6 questions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



Introduction

In the rapidly evolving field of data engineering, one constant remains: the ubiquitous presence of Linux. As data engineers, we often find ourselves interacting with Linux-based systems, whether it's managing databases, orchestrating data pipelines, or deploying machine learning models. However, a common challenge that many data engineers face, especially those at level 5, is the lack of comprehensive understanding and proficiency in Linux. This gap in knowledge can lead to inefficiencies and roadblocks in their workflow, hindering their ability to fully leverage the power of their data infrastructure.

Understanding the relevance of Linux for data engineers is crucial. Linux offers a robust and flexible environment that is ideal for handling large datasets and complex computations. It provides a wide range of tools and utilities that can be used to manipulate and process data efficiently. Moreover, Linux's open-source nature allows for high customisability, making it adaptable to various data engineering

needs. For level 5 data engineering students, mastering Linux can significantly enhance their skill set, making them more versatile and effective in their roles.

Consider the case of a real-world data engineering project at a leading tech company. The data engineering team was tasked with setting up a reliable and scalable data pipeline to handle the company's growing data needs. The team chose to use a Linux-based system due to its stability, performance, and the wide array of data processing tools available. However, the team's limited familiarity with Linux led to bottlenecks in the project, with tasks taking longer than necessary due to the learning curve associated with the Linux environment. This case underscores the importance of having a solid understanding of Linux in data engineering roles.

By the end of this topic, you will have gained a deep understanding of Linux from a data engineering perspective. You will learn how to navigate the Linux filesystem, use key command-line tools, write shell scripts to automate tasks, manage processes and resources, and secure a Linux system. These skills will not only make you a more efficient data engineer but also open up new opportunities for you. With a strong foundation in Linux, you will be able to design and manage more robust data infrastructures, contribute to open-source data projects, and stay ahead in the ever-evolving field of data engineering.

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: Setting up and getting started with NDG Linux Unhatched

NDG Linux Unhatched allows students to wade into the shallow end of Linux, the back-end operating system used by global titans such as Facebook, Google, Microsoft, NASA, Tesla, Amazon and more.

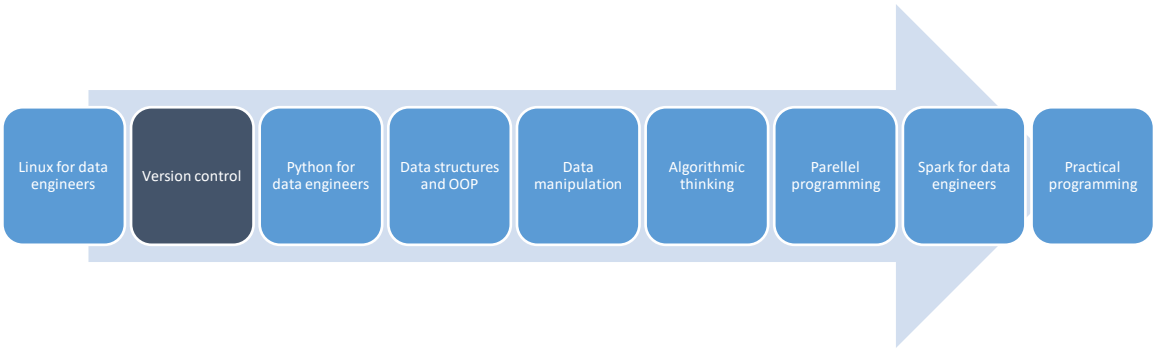
The aim of this task is to not only enable you to learn new concepts but also apply them practically, setting a strong foundation for your future learning.

To ensure you don’t get stuck, you will be guided step-by-step through a series of hands-on virtual machine activities.

Further guidance can be found here: L5DE 3.1 Apply (OTJ)

Link

This 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 **Version control**.