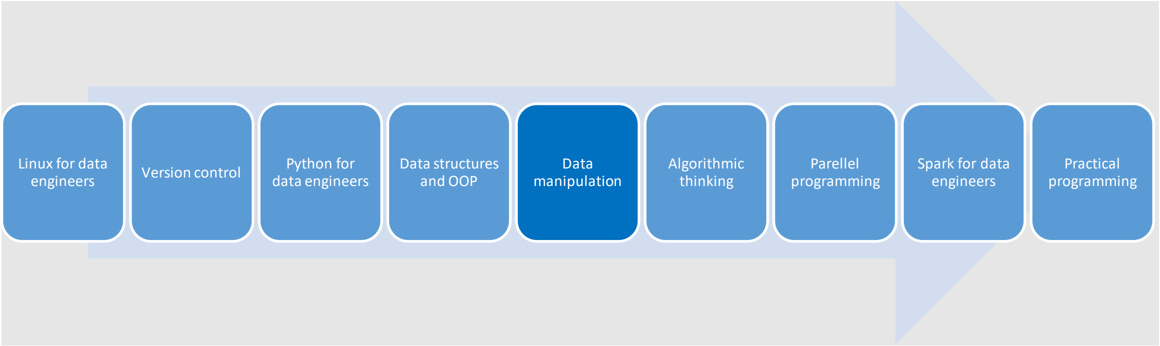
Topic 5 – Data manipulation

This document is the handbook for Topic 5 – **Data manipulation** – 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:

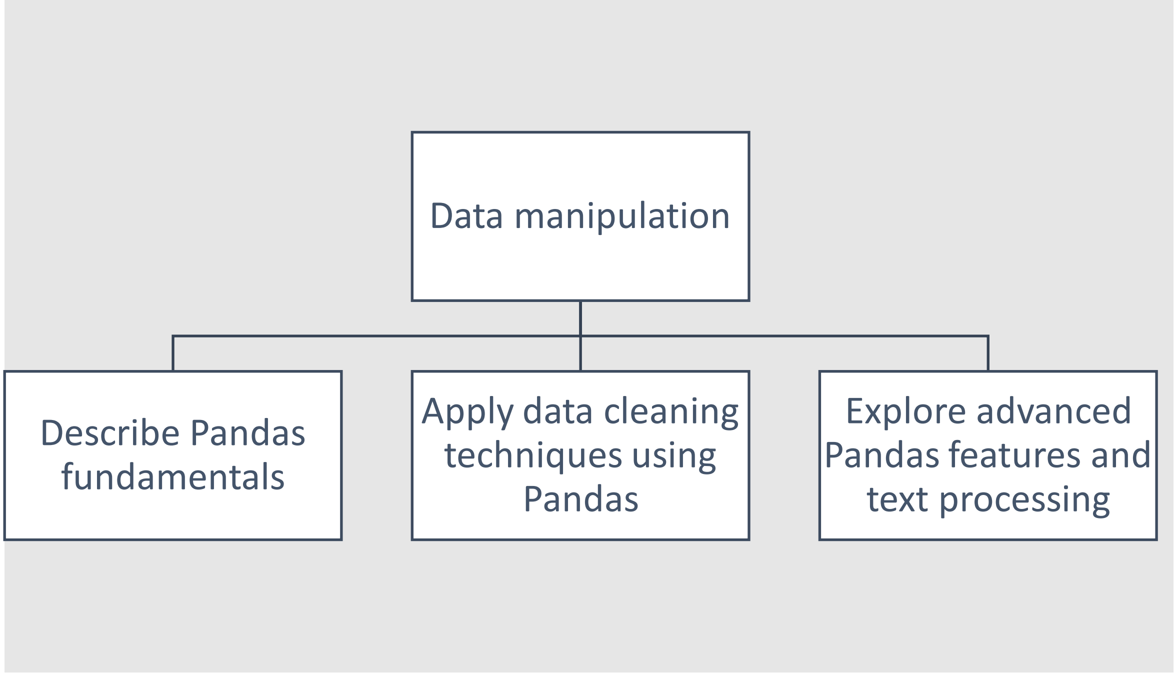
- Learners will have 10 weeks to complete the project and submit the report to the EPAO
- Learners also need to prepare and give a presentation to an independent assessor on their project
- The presentation with questions will last at least 50 minutes. The independent assessor will ask at least 6 questions 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 manipulation
Describe Pandas fundamentals
Apply data cleaning techniques using Pandas
Explore advanced Pandas features and text processing

Introduction

In the modern era, we are surrounded by an overwhelming amount of information. However, the true value of this information does not lie in its sheer volume, but in how we utilise it. Consider a vast library filled with countless books. Without the ability to organise, search, or extract wisdom from these books, the library's potential remains untapped. This is where data manipulation comes into play, serving as a powerful tool that allows us to transform raw data into actionable insights.

As individuals aspiring to become data engineers, we are faced with the challenge of dealing with a wide variety of data. This data, arriving in diverse formats such as spreadsheets, databases, logs, APIs, and more, can often seem like an unruly beast that needs to be tamed.

For instance, imagine a large retail company inundated with sales records. These records, amounting to millions of rows, are scattered across various files. The challenge lies in extracting meaningful trends and identifying top-selling products from this chaotic sea of information.

Data manipulation is crucial in these scenarios as it helps unlock hidden patterns within raw data. By cleaning, transforming, and aggregating data, we can unearth valuable insights that were previously hidden. These insights fuel decision-making processes within organisations, whether it’s optimising supply chains, predicting stock prices, or tailoring marketing campaigns. Moreover, mastering data manipulation can accelerate one’s career, as employers are always on the lookout for data engineers who can skilfully handle data. This mastery can open doors to exciting roles across various sectors such as tech, finance, healthcare, and more.

In terms of skills, this topic will focus on Python, SQL, and Pandas. We will use Python to craft scripts that can reshape data, filter rows, and merge tables seamlessly. SQL, or Structured Query Language, will be our tool for querying databases, joining tables, and extracting insights. Pandas, a versatile tool, will be used to slice, dice, and transform data effortlessly.

In conclusion, data manipulation is not just about writing code. It’s about unravelling the stories hidden within the intricate tapestry of data. So, let’s grab our keyboards, put on our data capes, and dive into the fascinating world of data manipulation.

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.

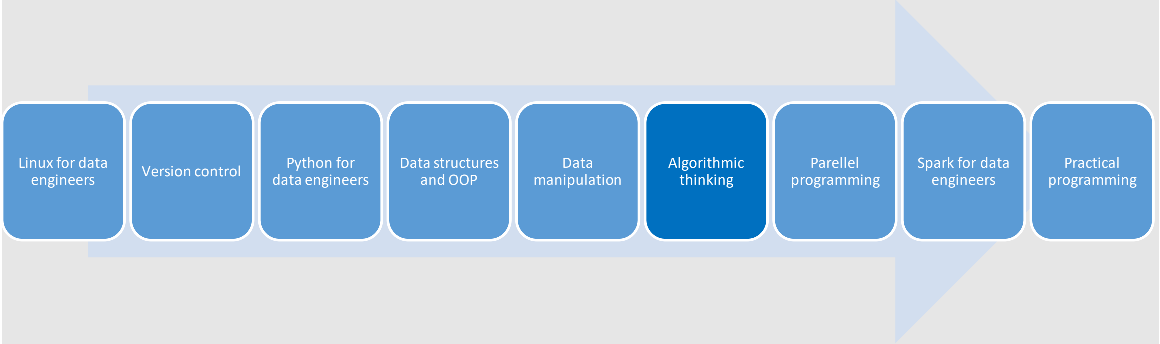
Off-the-job learning tasks

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: Your tutor will provide you with a brief on how to complete the Python notebook for this topic.

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 **Parallel programming**.