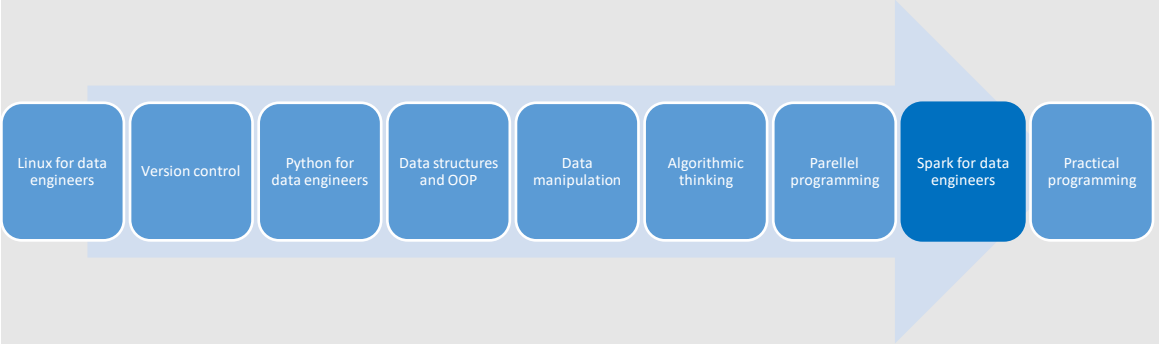
Topic 8 – Spark for data engineers

This document is the handbook for Topic 8 –**Spark 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:

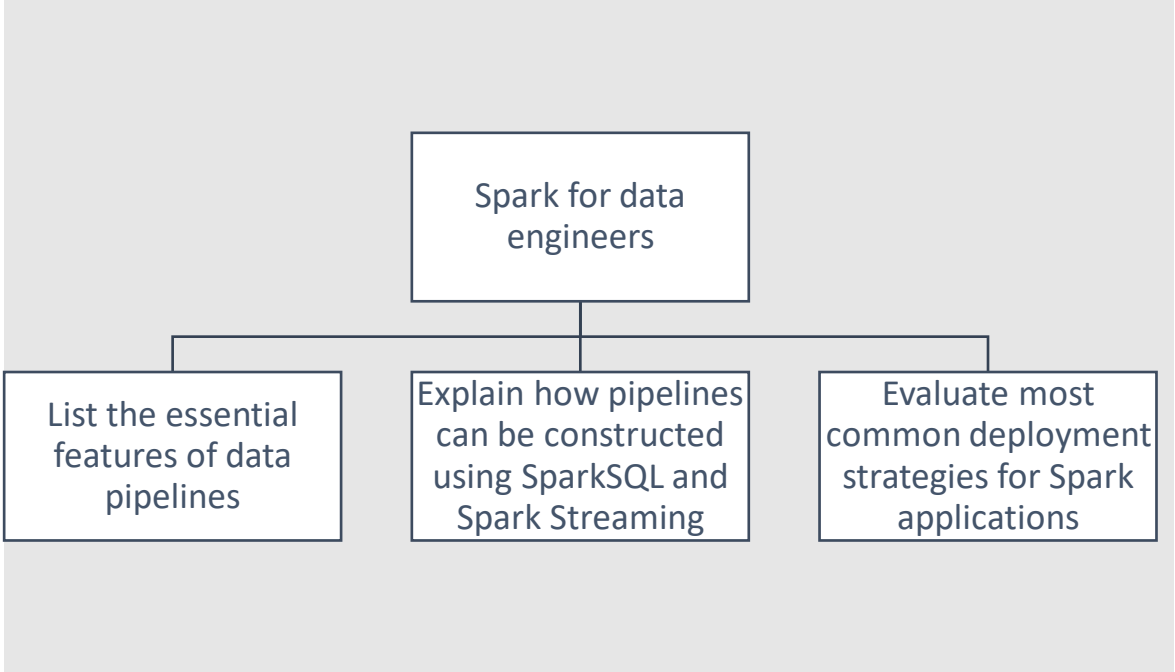
- Learners will have 10 weeks to complete the project and submit the report to the EPAO
- Learners also need to prepare and give a presentation to an independent assessor on their project
- The presentation with questions will last at least 50 minutes. The independent assessor will ask at least 6 questions 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 era of big data, the ability to process and analyse large datasets quickly and efficiently is crucial. Apache Spark, with its ability to handle large-scale data processing tasks, has emerged as a leading tool in this space. This topic is designed to equip you with the knowledge and skills to harness the power of Spark for data processing, and to construct efficient data pipelines using SparkSQL and Spark Streaming.

Data pipelines are essential for the efficient and reliable processing of data. They allow for the automation of data flow between sources and destinations, and they ensure that data is cleaned, transformed, and stored in a way that is ready for analysis. With SparkSQL and Spark Streaming, data engineers can construct robust data pipelines that can handle both batch and real-time data processing.

SparkSQL allows for the processing of structured and semi-structured data using SQL-like syntax, making it accessible for those familiar with SQL. It also integrates seamlessly with the Spark ecosystem, allowing for the use of other Spark libraries in conjunction with SparkSQL. On the other hand, Spark Streaming enables the processing of live data streams in real-time, making it ideal for applications that require immediate insights from data.

When it comes to deploying Spark applications, there are several strategies that data engineers can evaluate. These include standalone deployment, where Spark runs on its own cluster; on-premise deployment, where Spark runs on a company's internal hardware; and cloud-based deployment, where Spark runs on a cloud platform like AWS or Azure. Each of these deployment strategies has its own advantages and considerations, and the choice of strategy would depend on the specific requirements of the application.

By the end of this topic, you will be able to evaluate the use of Spark clusters for data processing, understand the essential features of data pipelines, and construct your own using SparkSQL and Spark Streaming. You will also have a deeper understanding of how data engineers with Spark skills can deliver real-world value.

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

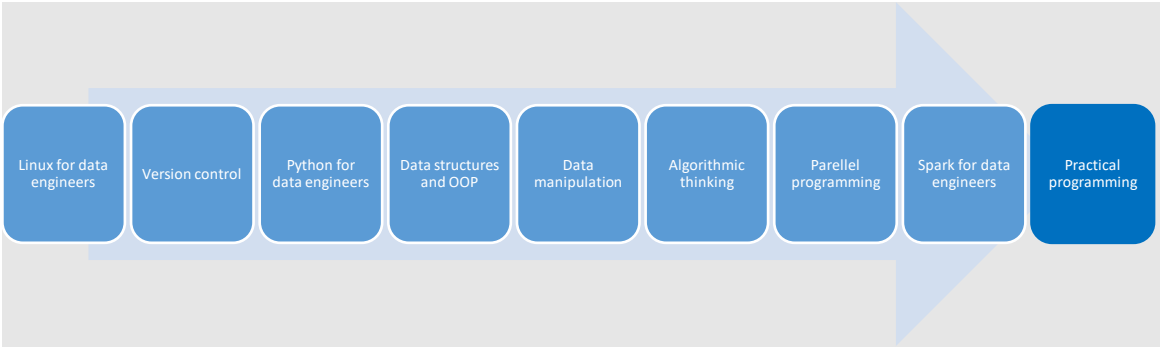
This stage is all about ensuring you truly grasp and retain what you've learned. Through completion of off-the-job (OTJ) revision tasks and tests, you'll get plenty of practice applying your knowledge. Plan to dedicate 6-8 hours each week to guided study and portfolio work, with sessions typically on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