Module 6 Topic 2

Worksheet

Portfolio Piece 3 – Mini Project – Guidance

Your third portfolio piece is a mini project containing one large piece of evidence, such as an architecture diagram. The word count is 750 words + a diagram.

The aim of the formative submissions of the portfolio (spread throughout the programme) is for you to build enough evidence that you have mastered the learning outcomes for your programme. You may then find it useful to refer to pieces from your portfolio evidence in your professional discussion during the EPA (final) stage of the programme. We will be providing feedback on your portfolio submissions. There will be no formal marks or grades.

The mini project allows you to demonstrate a comprehensive understanding of data ingestion architectures and to apply this knowledge in a practical and coherent manner. Choose a realistic scenario that involves data collection and ingestion. This could be a retail company looking to improve inventory management through real-time analytics, a financial institution needing to process large volumes of transaction data, or a healthcare provider aiming to aggregate patient data from multiple sources. It could be a use case from your own organisation (make sure you discuss it with your manager and ensure you are clear on confidentiality considerations, if they apply).

Develop an Architecture Diagram:
Create a detailed diagram that illustrates the data ingestion pipeline. Include components such as data sources, landing, staging, and ingestion layers, as well as the data storage solutions (e.g. NoSQL databases). Highlight the flow of data through the system, showing how data moves from one component to another. Indicate mechanisms for data validation, cleaning, and integration.

Write an Explanation:
In 750 words, describe the architecture and explain how each component of the diagram functions within the overall system. Discuss the choices of technology and the rationale behind the design decisions, including considerations of scalability, reliability, and security. Reflect on how this architecture helps achieve the business objectives outlined in your chosen scenario.

Feedback and Revision
Utilise the feedback provided on this portfolio piece to refine your understanding and presentation of data ingestion architectures. Consider how you can improve the clarity of your diagrams and the effectiveness of your written explanations. Prepare

to discuss this project during your EPA stage, focusing on explaining your technical choices and how you have addressed potential challenges in your design. By engaging deeply with this project, you will not only demonstrate your technical capabilities but also develop a robust understanding of the practical applications of data ingestion architectures in various industries. This experience will be instrumental in your progression as a Level 5 Data Engineer.

Mini-Project Requirements

Scenario Selection and System Design:
Choose a data-intensive scenario from sectors like retail, finance, or healthcare, or your industry.

Create an architecture diagram that reflects a heterogeneous data ingestion system capable of handling multiple data formats and sources.

Automation in Data Collection and Ingestion (Learning Outcome 1):
Justify the use of automation in your architecture, discussing its impact on efficiency and accuracy in data collection and ingestion processes.

Data Cleaning Techniques (Learning Outcome 2):
Evaluate common data cleaning techniques applicable in your designed system. Describe how these techniques integrate into the staging area of your data pipeline to ensure data quality.

Pre-processing for Machine Learning (Learning Outcome 3):
Identify and outline the steps required to preprocess data within your system specifically for machine learning applications, highlighting any automated transformations or feature engineering tasks.

Practical Data Collection Skills (Learning Outcome 4):
Demonstrate these skills through your detailed explanation of the data collection mechanisms in your diagram, including specifics on how data is extracted from diverse sources.

Design and Architecture Considerations (Learning Outcome 5):
Explain the key design and architecture considerations in building a heterogeneous data ingestion system, focusing on scalability, reliability, security, and compliance.

Usefulness of Heterogeneous Ingestion Patterns (Learning Outcome 6):
Report on how the heterogeneous ingestion patterns utilised in your design can address specific business use cases, enhancing operational efficiency and decision-making.

Practical Application of File Ingestion (Learning Outcome 7):
Illustrate through your architecture how multiple sources of data are combined effectively in the ingestion process, showing practical application skills in handling files from disparate sources.

BPP