Due: (End of Module 3)

Brief for Level 5 Data Engineer - Portfolio Piece 1

"Data Quality and Performance"

Word count: up to 1,500 words (+/- 10%)

Objective: *Highlight how the business achieves a higher level of performance through the use of good quality data.*

Deliverables: *Submit **at least one item of evidence per section** to support your work, including **screenshots** and brief explanations for each. Submit via the Hub by the deadline at the top of the document.*

Outline:

As a data engineer, you are responsible for ensuring that the systems and services you manage are robust, efficient, and trustworthy. This activity asks you to reflect on how you monitor and manage data systems, address data quality challenges, and use various analytics and ingestion strategies to maintain performance.

Write a professional report that addresses each of the following areas. Use real examples from your work or training to illustrate your understanding and practice.

Submission Requirements:

- Submit your work in a single PDF or Word document.
- Caption and number each screenshot (For example: Fig. 1 - My important image) and then refer to your numbered screenshots to explain the actions taken (for example "As seen in Fig. 1, I have...").
- Save this document both in your learning journal and on the Hub.

Tasks

1. Monitoring Data Stores for Performance and Availability

Explain the different types of data stores you work with (e.g. SQL, NoSQL, distributed file systems) and how you monitor them. Describe:

- What tools you use (e.g. your cloud-based or on-premise provider, along with specific storage solutions and the data in them).
- How monitoring helps you optimise system management, ensure high availability, and maintain performance.

Compare and contrast the different types of data stores you have used, focusing on how their choice for the data stored have helped to optimise their use.

2. Data Normalisation and Its Advantages

Define the principles of data normalisation. Explain how it helps to:

- Reduce data redundancy.
- Protect against inconsistent dependencies.
- Contribute to overall data integrity and quality.

Use an example from your business of when data has been normalised and how the above points have shown benefits.

3. Data Quality and Risk Management

Discuss the inherent risks in working with data, including (but not limited to):

- Incomplete or inaccurate data.
- Use of unethical or non-compliant data sources.
- Risks from data drift, human error, or integration failures.

Explain your approach to ensuring data quality. Include methods such as validation rules, data profiling, data cleansing, and audit trails. Reflect on why these practices are essential for building trustworthy data pipelines.

4. Data Ingestion Frameworks and Optimisation

Describe how you use or design data ingestion processes, considering:

- The use of batch, streaming, and on-demand services.
- Why you would choose one method over another.
- How these methods impact performance, timeliness, and scalability.

Include an example where you moved data from one location to another (e.g. from a cloud store to a data warehouse) and how your choice of ingestion method improved efficiency or reliability.

Extension: Descriptive, Predictive and Prescriptive Analytics

While you may not have covered this necessarily as part of your learnings so far, you might want to read these articles:

4 Types of Data Analytics to Improve Decision-Making

Types of Data Analytics in Engineering

Explain the differences between **descriptive**, **predictive**, and **prescriptive** analytics:

- What each one aims to achieve.
- Typical techniques involved (e.g. visualisation, machine learning, optimisation algorithms).
- How they support decision-making in a business context.

Provide an example where you've used (or could use) each approach to extract value from data.

Guidelines for Writing

Structure

- It is suggested here that you take around 300 words per section, with a short introduction to highlight your role and responsibilities (keep this brief). You don't need to hit the limit - focus on being clear and well-explained.
- Split your work into **five clear sections**, one for each topic.
- Use simple **headings** (like "Monitoring Data Stores") to keep your writing organised.

Tone and Style

- Write in a **professional but friendly** tone, as if you're explaining your thinking to a team member or line manager.
- Use **real examples** from work, projects, or learning - these help show what you understand.
- Bullet points are fine to explain technical ideas but try to include **full sentences and short paragraphs** to connect your thoughts.

Evidence of Understanding

- Try to **explain terms in your own words**. If you mention a tool or method, briefly say how you use it or why it matters.
- Where possible, give **short scenarios** to demonstrate your thinking - like how you'd use streaming in a real data pipeline.

Tips for Success

- Start each section by showing what you understand.
- Then give an example or situation.
- Finally, explain why that it is useful or important in your role as a data engineer.

Assessment Criteria:

- Shared at least one unique example, scenario, or experience for each section to show how to apply your knowledge.
- Shown an understanding of the assessment by explaining key concepts in your own words.
- Described why these ideas are important in a data engineering context.
- Work is easy to follow, well structured, and professionally presented.
- Used headings, simple explanations, and connected ideas together.

KSBs

These are the KSBs you will be focusing on as part of this assessment.

KSB	Descriptor
K1	Processes to monitor and optimise the performance of the availability, management and performance of data product.
K3	Data normalisation principles and the advantages they achieve in databases for data protection, redundancy, and inconsistent dependency.
K5	The inherent risks of data such as incomplete data, ethical data sources and how to ensure data quality.
K18	How to use streaming, batching and on-demand services to move data from one location to another.
K27	The principles of descriptive, predictive and prescriptive analytics.
S7	Work with different types of data stores, such as SQL, NoSQL, and distributed file system.
S15	Optimise data ingestion processes by making use of appropriate data ingestion frameworks such as batch, streaming and on-demand.