INTERNATIONAL QUALIFICATIONS

AND ASSESSMENT CENTRE (IQAC)

«Approved»

2025 y. «16» 04

Programme	**Level 6 Diploma in Data Science**		
Unit Number/ Unit Title	UNIT 4 DATA ENGINEERING AND BIG DATA TECHNOLOGIES		
Cohort Code:	L06DEBD-U4		
Unit Level	Level 6		
Total GLH	Total qualification time 200/ Total Guided learning hours 90/ Self-guided learning hours 110		
Credits	20 CATS/ 10 ECTS		
Lecturer			
Start Date		**End Date**	

Unit Aims	This unit provides learners with practical knowledge of data engineering tools such as Apache Hadoop, Spark, Kafka, and modern ETL pipelines. Students will gain practical knowledge of essential data engineering tools, including Apache Hadoop, Spark, and Kafka, along with modern ETL (Extract, Transform, Load) pipelines. The unit will focus on hands-on experience in utilizing these technologies to manage and process large datasets efficiently. By the end of the course, students will be adept at implementing data workflows and optimizing data processing tasks, preparing them for real-world data engineering challenges.
Differentiation Strategies *(e.g. planned activities or support for individual learners according to their needs)*	The total number of students to be in the lesson is approximately 20. This is a multicultural group of students predominantly between the ages of 24 – 45, with numerous ethnic, gender, and creed background. These are UK academic level 5 students; hence it is assumed that they have practical, theoretical, or technological knowledge and understanding of a subject or field of work to find ways forward in broadly defined, complex contexts. These students must be able to generate information, evaluate, synthesise the use information from a variety of

	sources. Various approaches to addressing the various identified students needs will be adopted throughout the lesson. Such will include:- 1. Progressive tasks 2. Digital resources 3. Verbal support 4. Variable outcomes 5. Collaborative learning 6. Ongoing assessment 7. Flexible-pace learning
Equality & Diversity	Variety of teaching techniques will be employed to ensure that the needs of each individual learner are met.
Safeguarding & Prevent	Safeguarding policies and the Prevent duty are strictly observed to ensure the safety, well-being, and inclusivity of all students and staff.
Health & Safety	SIRM H&S policies will be maintained.
Learning Resources	**Teaching and Learning Materials** • White, T. (2015). Hadoop: The Definitive Guide. O'Reilly Media. • Karau, H., & Warren, R. (2017). High Performance Spark. O'Reilly Media. • Guller, M. (2020). Big Data Analytics with Spark. Apress.

Learning Outcome	Assessment Criteria
LO1. 1. Understand the architecture of big data systems.	Written Report: 1.1 Explain batch vs stream processing. 1.2 Describe the Hadoop ecosystem.
LO2. 2. Use ETL tools to process large-scale datasets.	Programming Task: 2.1 Build ETL pipelines using Apache NiFi or Airflow. 2.2 Transform data using Spark SQL or PySpark.
LO3. 3. Apply streaming technologies for real time analytics.	Lab Work: 3.1 Use Apache Kafka or Flink. 3.2 Monitor data streams for processing efficiency.
LO4. 4. Ensure data quality and system reliability.	Portfolio: 4.1 Design validation rules for pipelines. 4.2 Address fault tolerance in distributed environments.

No	Learning Outcome / Topic	Learning and Teaching Activities	Which assessment criteria does the session relate to?	Day/month/year/ signature
1.	**Introduction to Big Data**	**Introduction to Big Data** 5 V's (Volume, Velocity, Variety, Veracity, Value)	LO1: Big Data System Architecture	
2.	**Batch vs. Stream Processing**	**Batch vs. Stream Processing** Lambda vs. Kappa architectures	LO1: Big Data System Architecture	
3.	**Hadoop Ecosystem**	**Hadoop Ecosystem** HDFS, YARN, MapReduce	LO1: Big Data System Architecture	
4.	**Modern Data Stack**	**Modern Data Stack** Data lakes (Delta Lake, Iceberg), lakehouses	LO1: Big Data System Architecture	
5.	**Cloud vs. On-Prem Solutions**	**Cloud vs. On-Prem Solutions** AWS EMR, Databricks, Cloudera	LO1: Big Data System Architecture	
6.	**ETL Fundamentals**	**ETL Fundamentals** Extract, Transform, Load workflows	LO2: ETL & Data Processing	
7.	**Batch ETL Tools**	**Batch ETL Tools** Apache Spark, AWS Glue, Talend	LO2: ETL & Data Processing	
8.	Half-Term Exam	- Review of LO1 topics - Practice questions and mock assessment - **Half-term assessment** based on LO1 (theory)	LO1 LO2	
9.	**Data Pipeline Orchestration**	**Data Pipeline Orchestration** Airflow, Prefect, Dagster	LO2: ETL & Data Processing	

10.	**Data Transformation Techniques**	**Data Transformation Techniques** SQL-based (dbt), PySpark	LO2: ETL & Data Processing	
11.	**Incremental Data Loading**	**Incremental Data Loading** CDC (Change Data Capture), SCD (Slowly Changing Dimensions)	LO2: ETL & Data Processing	
12.	**Stream Processing Basics**	**Stream Processing Basics** Event time vs. processing time, watermarks	LO3: Streaming & Real-Time Analytics	
13.	**Apache Kafka**	**Apache Kafka** Topics, partitions, producers/consumers	LO3: Streaming & Real-Time Analytics	
14.	Final Exam Preparation & Review	- Comprehensive review of all learning outcomes - Practice questions and revision of key topics		
15.	Final Exam	- **Final-term assessment** covering all learning outcomes (theory and practical elements)		
16.	Feedback & Reflection	- Review of final exam - Individual feedback on performance - Reflective discussion on key learning points		
17.	**Stream Processing Frameworks**	**Stream Processing Frameworks** Apache Flink, Spark Streaming	LO3: Streaming & Real-Time Analytics	
18.	**Real-Time Analytics**	**Real-Time Analytics** Windowed aggregations, stateful processing	LO3: Streaming & Real-Time Analytics	

19.	**Use Cases**	**Use Cases** Fraud detection, IoT monitoring	LO3: Streaming & Real-Time Analytics	
20.	**Data Quality Dimensions**	**Data Quality Dimensions** Accuracy, completeness, consistency	LO4: Data Quality & Reliability	
21.	**Data Validation Tools**	**Data Validation Tools** Great Expectations, Deequ	LO4: Data Quality & Reliability	
22.	**Monitoring & Alerting**	**Monitoring & Alerting** Prometheus, Grafana, custom dashboards	LO4: Data Quality & Reliability	
23.	Half-Term Exam	**Project** End-to-end data pipeline with quality checks		
24.	**Error Handling & Recovery**	**Error Handling & Recovery** Dead-letter queues, retry mechanisms	LO4: Data Quality & Reliability	
25.	**Data Governance**	**Data Governance** Metadata management, lineage tracking	LO4: Data Quality & Reliability	
26.	**Modern Data Platforms**	**Modern Data Platforms** Snowflake, BigQuery, Redshift	LO5: Capstone & Emerging Trends	
27.	**Data Mesh Principles**	**Data Mesh Principles** Domain-oriented ownership	LO5: Capstone & Emerging Trends	
28.	**ML Data Pipelines** and **Ethical Considerations**	**ML Data Pipelines** Feature stores (Feast, Tecton) **Ethical Considerations** Bias in big data, privacy preservation	LO5: Capstone & Emerging Trends	
29.	Final Exam Preparation & Review	LO1, LO2, LO3, LO4	LO1, LO2, LO3, LO4	
30.	Final Exam		LO1, LO2, LO3, LO4	