



GANPAT UNIVERSITY

FACULTY OF COMPUTER APPLICATIONS

Programme	Master of Computer Applications					Branch/Spec.	Computer Application		
Semester	II					Version	1.0.0.0		
Effective from Academic Year			2026-27			Effective for the batch Admitted in		July 2026	
Subject Code	P12E6DPE		Subject Name			Data Pipelines & ETL Orchestration			
Teaching scheme						Examination scheme (Marks)			
(Per week)	Lecture (DT)		Practical (Lab.)		Total		CIA	SEE	Total
	L	TU	P	TW					
Credit	2	0	2	0	4	Theory	40	60	100
Hours	2	0	4	0	6	Practical	20	30	50

Objective:

The objective of this course is to provide students with the knowledge and practical skills required to design, develop, automate, and manage modern data pipelines for enterprise applications. Students will learn the principles of data engineering, ETL/ELT processes, workflow orchestration, and real-time streaming technologies to build scalable, reliable, and secure data integration solutions using industry-standard tools and best practices.

Pre-requisites:

Students should have basic knowledge of:

- Relational Database Management Systems (RDBMS)
- SQL and Database Queries
- Fundamentals of Programming (Python preferred)
- Operating Systems and Linux basics
- Basic understanding of Data Structures and Computer Networks

Course Outcomes:

- 1 = Slight (Low); 2 = Moderate (Medium); 3 = Substantial (High); “-” = No Correlation

Name of CO	Description
CO1	Explain the fundamentals of data engineering, modern data pipeline architectures, data storage platforms, and enterprise data governance principles.
CO2	Design and develop ETL/ELT pipelines by applying data extraction, transformation, loading, data quality, and integration techniques.
CO3	Build, automate, schedule, monitor, and optimize enterprise data pipelines using workflow orchestration and CI/CD concepts.
CO4	Develop real-time streaming data pipelines using Apache Kafka and event-driven architectures for scalable data processing applications.
Mapping of CO and PO	

Cos	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8
CO1	3	2	1	2	1	-	2	1
CO2	3	3	3	3	2	1	2	2
CO3	2	3	3	3	3	2	2	2
CO4	2	3	3	3	2	2	2	3

CO-PO Reference Link : <https://ampics.guni.ac.in/programmes/after-graduation-pg-programs/computer-applications/master-of-computer-applications>

Content:

Unit	SECTION-I	Hours
01	Foundations of Data Pipelines and Data Engineering: Introduction to Data Engineering, roles and responsibilities of Data Engineers, Data Engineering lifecycle, modern data pipeline architecture, batch and real-time pipelines, enterprise data platforms, operational databases, data warehouse, data lake, data lakehouse, data mesh concepts, relational database fundamentals, dimensional modeling, star and snowflake schemas, metadata management, master data management, data lineage, data governance, data privacy, security and enterprise data governance practices.	7
02	ETL, ELT and Data Pipeline Development: ETL and ELT architecture, ETL versus ELT, enterprise data integration lifecycle, data extraction from databases, APIs and files, data transformation techniques including cleansing, validation, standardization, enrichment and aggregation, business rule implementation, data loading strategies, full and incremental loading, Change Data Capture (CDC), Slowly Changing Dimensions (SCD), data quality management, data profiling, exception handling, enterprise data integration patterns and pipeline testing.	8
SECTION-II		
03	Data Pipeline Orchestration and Automation: Data pipeline architecture, pipeline design patterns, batch and event-driven workflows, workflow orchestration concepts, Apache Airflow architecture, Directed Acyclic Graphs (DAGs), operators, sensors, scheduling, workflow automation, task dependency management, event-based scheduling, monitoring, logging, alerting, retry mechanisms, fault tolerance, performance optimization CI/CD concepts and deployment best practices for enterprise data pipelines.	8
04	Streaming Data Pipelines and Data Engineering: Stream processing fundamentals, event-driven architecture, real-time data processing, Apache Kafka architecture, topics, partitions, producers, consumers, Kafka Connect and Kafka Streams.	7

Practical Content:

List of programs specified by the subject teacher based on above mentioned topics.

Text Books:

1	Joe Reis and Matt Housley, <i>Fundamentals of Data Engineering: Plan and Build Robust Data Systems</i> , O'Reilly Media, First Edition, 2022.
2	Martin Kleppmann, <i>Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems</i> , O'Reilly Media, First Edition, 2017.
3	Ralph Kimball and Joe Caserta, <i>The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data</i> , Wiley, First Edition, 2004
4	Tyler Akidau, Slava Chernyak, and Reuven Lax, <i>Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing</i> , O'Reilly Media, First Edition, 2018
5	Neha Narkhede, Gwen Shapira & Todd Palino, <i>Kafka: The Definitive Guide</i> , 2nd Edition, O'Reilly Media.

Reference Books:

1	Martin Kleppmann, <i>Designing Data-Intensive Applications</i> , O'Reilly Media.
2	Ralph Kimball & Margy Ross, <i>The Data Warehouse Toolkit</i> , 3rd Edition, Wiley.
3	Bill Inmon, <i>Building the Data Warehouse</i> , 5th Edition, Wiley.
4	Maxime Beauchemin & Julien Le Dem, <i>The Definitive Guide to Apache Airflow</i> , O'Reilly Media.
5	Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia, <i>Learning Spark</i> , 2nd Edition, O'Reilly Media.

MOOC/Certification Courses:

1	Coursera (IBM)- IBM Data Engineering Professional Certificate
2	Coursera (Google Cloud) - Preparing for Google Cloud Professional Data Engineer
3	Udemy- Apache Kafka for Beginners using Hands-on Labs

Question Paper Scheme:

University Examination Duration: 3 Hours

Note for Examiner: -

(I) Questions 1 and 4 are compulsory with no options.

(II) Internal options should be given in questions 2, 3, 5 and 6.

SECTION – I

Q.1 –8 Marks

Q.2 –10 Marks

Q.3 –12 Marks

SECTION - II

Q.4 –8 Marks

Q.5 –10 Marks

Q.6 –12 Marks