(Affiliated to University of Mumbai)	***End Semester Examination (R-24/R-25) SH 2026***			
Branch: AIDS	Course: Data Engineering			
Year/ Semester: TE/V	Course code: AIDSC502			
Time: 03 hours	Marks: 80			
Note: 1. All questions are compulsory. **2. Figures to right indicate full marks.** **3. Assume suitable data wherever necessary.**		Marks	CO	BL
Q. 1	All questions are compulsory.	**30**		
A.	Describe the five stages of the data engineering lifecycle for a food delivery company processing order events and GPS updates. Give one suitable activity for each stage to support raw data storage, delivery-time calculation, and operational dashboards. Solution: The five stages are generation, ingestion, storage, transformation, and serving. Activities include collecting order/GPS events, ingesting them, storing raw data, calculating delivery times, and presenting results on operational dashboards. Marking scheme: Identification of five stages – 2.5 marks; suitable activity for each stage – 2.5 marks.	05	CO1	L2
B.	Differentiate file, block, and object storage based on their structure and scalability. Select the suitable storage type for departmental documents, transactional databases, product images, and application logs in an e-commerce company, with justification. Solution: File storage suits departmental documents, block storage suits transactional database volumes, and object storage suits product images and application logs because it supports scalable storage of large volumes of unstructured data. Marking scheme: Storage types and their features – 2 marks; suitable selection for the four applications – 2 marks; justification – 1 mark.	05	CO2	L2
C.	Illustrate the differences between an Internal Data Engineer and External Data Engineer in terms of their roles, responsibilities, and key functions, with suitable examples. Solution: Internal Data Engineers work within an organization to build and maintain its data infrastructure, while External Data Engineers work for external clients or service providers to develop, integrate, or manage data solutions. Marking scheme: Roles and responsibilities of Internal Data Engineers – 2 marks; roles and responsibilities of External Data Engineers – 2 marks; suitable examples – 1 mark.	05	CO1	L3
D.	Analyze the key features of data lineage and examine how they help track data origins, transformations, and dependencies in a data pipeline.	05	CO3	L4

	Solution: Data lineage tracks data origins, transformations, and dependencies, helping identify errors, trace data quality issues, and understand how changes in source data affect downstream outputs. Marking scheme: Key features of data lineage – 2 marks; tracking origins, transformations, and dependencies – 2 marks; relevance to data quality and error tracing – 1 mark.			
E.	Examine the stages of the SQL query lifecycle and determine their roles using a neat diagram. Solution: The stages are query submission, parsing and validation, query optimization, execution, and result retrieval; each stage helps process the SQL query and return the required results. Marking scheme: Identification of lifecycle stages – 2.5 marks; explanation of their roles using a neat diagram – 2.5 marks.	05	CO4	L3
F.	A university stores student records, examination results, and fee details in its database and transmits information through its student portal. Elaborate on how encryption at rest and encryption in transit protect sensitive student data, with suitable examples. Solution: Encryption at rest protects stored student records, examination results, and fee details, while encryption in transit (e.g., HTTPS/TLS) protects information transmitted between the student portal and server. Marking scheme: Encryption at rest with an example – 2 marks; encryption in transit with an example – 2 marks; relevance to student data protection – 1 mark.	05	CO6	L4
Q.2A	A healthcare organization collects structured patient records, semi-structured medical reports, and unstructured medical images from multiple hospitals. Illustrate, using a neat diagram, how a data lake architecture stores, manages, and processes these diverse data types for healthcare analytics. Solution: A data lake stores structured patient records, semi-structured medical reports, and unstructured medical images in raw form. Its layers support data ingestion, metadata management, processing, and analytics to enable healthcare insights. Marking scheme: Data lake architecture diagram with major components – 3 marks; storage and handling of the three data types – 3 marks; data management and processing – 2 marks; healthcare analytics applications – 2 marks.	10	CO2	L3
	OR			
Q.2B.	A company uses BigQuery to analyze streaming order data and stores historical data in object storage. Apply the concepts of compute–storage separation and hybrid object storage to explain how the company can manage real-time and historical data efficiently.	10	CO2	L3

	Solution: Compute–storage separation allows BigQuery to scale query-processing resources independently of storage, enabling efficient analysis of streaming orders. Hybrid object storage keeps historical data in scalable, low-cost storage while allowing analytical queries across current and historical data. Marking scheme: Compute–storage separation and its benefits – 3 marks; hybrid object storage and historical data management – 3 marks; application to real-time order analysis – 2 marks; efficiency, scalability, and cost benefits – 2 marks.			
Q.3A	Analyze how data frequency, bounded/unbounded data, and throughput affect real-time congestion predictions in smart city traffic system. Suggest suitable ingestion methods and improvements to reduce latency and improve real-time decision-making. Solution: High-frequency, unbounded traffic-sensor streams require real-time ingestion, while bounded historical traffic reports suit batch ingestion. High throughput requires scalable stream-processing systems; edge processing, parallel ingestion, and in-memory processing reduce latency and improve congestion predictions. Marking scheme: Impact of data frequency, bounded/unbounded data, and throughput – 3 marks; suitable ingestion methods – 3 marks; latency-reduction strategies – 2 marks; benefits for real-time congestion prediction and decision-making – 2 marks.	10	CO3	L4
	OR			
Q.3B	Analyze how data ingestion, push/pull communication, Time-to-Live (TTL), and Dead-Letter Queue (DLQ) mechanisms can be integrated to design a reliable real-time data streaming pipeline for an online food delivery platform. Explain how the pipeline handles delayed and duplicate events, processing failures, and data retention. Solution: Use Kafka or a similar platform to ingest real-time order updates. Push communication sends events immediately, while pull communication lets consumers retrieve events at their own pace. TTL manages event retention, and DLQ stores repeatedly failed messages. Event IDs, timestamps, retries, and idempotent processing handle duplicate and delayed events and ensure reliable recovery. Marking scheme (10 marks): Ingestion and communication – 3 marks; TTL and DLQ – 3 marks; error handling, duplicate management, and retention – 4 marks.	10	CO3	L4

Q.4A	A multinational retail company integrates customer, product, order, and payment data from multiple source systems into a centralized data warehouse. Construct a suitable Data Vault model by identifying the Hubs, Links, and Satellites required for this scenario. Illustrate how these components support data integration and historical data tracking. Solution: Create Hubs for Customer, Product, Order, and Payment to store unique business keys. Create Links such as Customer–Order, Order–Product, and Order–Payment to represent relationships. Add Satellites to store descriptive attributes and historical changes, including customer details, product descriptions, order status, and payment details. Load timestamps and source-system metadata support historical tracking, while Hubs, Links, and Satellites integrate data from multiple systems. Marking scheme (10 marks): Identification of Hubs – 2 marks; Links and relationships – 3 marks; Satellites and attributes – 2 marks; data integration and historical tracking – 3 marks.	10	CO4	L3
	OR			
Q.4B	An online retailer loads its daily sales file into a data warehouse every night at 1:00 AM. The steps are: check that the file has arrived, clean the data, load it into the warehouse, and send a summary email. i) Construct an Apache Airflow DAG for this workflow. Show the tasks, operators, dependencies and schedule. ii) Apply Airflow's monitoring and logging features to handle a failed load task. Solution: Create an Airflow DAG scheduled daily at 1:00 AM with four tasks: check_file using FileSensor, clean_data using PythonOperator, load_warehouse using PythonOperator, and send_email using EmailOperator. Set dependencies as check_file → clean_data → load_warehouse → send_email. Use Airflow's task logs and monitoring UI to identify load errors, configure retries and retry delays, and set failure alerts through email callbacks. Marking scheme (10 marks): DAG tasks and suitable operators – 3 marks; dependencies and schedule – 2 marks; monitoring and logging – 2 marks; failure handling using retries and alerts – 3 marks.	10	CO4	L3
Q.5A	Examine Apache Spark's architecture with a neat, labelled diagram. i)Describe the roles of the Driver Program, Cluster Manager and Executors, and explain how they interact during distributed processing. ii)Trace the execution of a Spark application from submission to completion, explaining how an action leads to jobs, stages and tasks, and how the results are returned or stored.	10	CO5	L3

	Solution: The Driver Program coordinates the Spark application, the Cluster Manager allocates resources, and Executors execute tasks and store intermediate data. When an action such as collect() or write() is invoked, Spark creates a job, divides it into stages based on shuffle boundaries, and splits stages into tasks. The Driver schedules tasks on Executors, which process data partitions in parallel. Results are returned to the Driver or written to storage. Marking scheme (10 marks): Labelled architecture diagram – 2 marks; roles and interaction of Driver, Cluster Manager, and Executors – 3 marks; execution flow from application submission to completion, including jobs, stages, and tasks – 3 marks; result handling or storage – 2 marks.			
	OR			
Q.5B	A manufacturing company uses shared-memory multiprocessors for production simulations and networked computers for quality analysis. Compare tightly coupled and loosely coupled systems, and recommend a suitable distributed-computing model for combining idle factory computers with dedicated computing resources. Solution: Tightly coupled systems use shared memory and a common operating system, enabling fast communication between processors. Loosely coupled systems consist of independent networked computers with separate memory and operating systems. A grid-computing model is suitable for combining idle factory computers with dedicated computing resources, as it coordinates distributed resources to perform quality-analysis tasks efficiently. Marking scheme (10 marks): Characteristics of tightly coupled systems – 2 marks; characteristics of loosely coupled systems – 2 marks; comparison – 2 marks; suitable distributed-computing model and justification – 4 marks.	10	CO5	L3
Q.6A	A hospital group stores patient records and appointment data on a public cloud platform accessed by engineers, analysts, and third-party billing vendors. Analyze the importance of data security and privacy, and examine how the principle of least privilege and the shared responsibility model can be applied to protect the data. Solution: Data security and privacy protect sensitive patient records from unauthorized access, data breaches, and misuse, ensuring confidentiality, integrity, and availability. The principle of least privilege grants engineers, analysts, and third-party billing vendors only the minimum access required for their roles. The shared responsibility model divides security duties between the cloud provider and hospital: the provider secures the underlying infrastructure, while the hospital manages access permissions, data protection, configurations, and compliance. Encryption, audit logs, and regular access reviews further strengthen protection.	10	CO6	L4

	Marking scheme (10 marks): Importance of data security and privacy – 3 marks; application of least privilege – 3 marks; shared responsibility model – 3 marks; additional security measures – 1 mark.			
	OR			
Q.6B	A retail group faces a data management bottleneck due to its centralized data team and plans to integrate machine learning into its data systems. Analyze how Data Mesh and ML-integrated data systems can address these challenges and support demand forecasting and anomaly detection. Solution: Data Mesh decentralizes data ownership by assigning retail domains, such as sales, inventory, and supply chain, responsibility for their data products. Self-service infrastructure and federated governance reduce dependence on the centralized data team. ML-integrated data systems use historical sales, seasonal trends, and inventory data for demand forecasting, while anomaly-detection models identify unusual transactions, sales fluctuations, and inventory discrepancies. Integrating ML into data pipelines enables automated predictions, timely alerts, and data-driven decisions. Marking scheme (10 marks): Data Mesh principles and decentralized ownership – 3 marks; ML integration into data systems – 2 marks; application to demand forecasting – 2 marks; anomaly detection and operational benefits – 3 marks.	10	CO6	L4
	*********** All the Best***********			