Daffodil International University

Faculty of Science & Information Technology
Department of Computing and Information System (CIS)
Final Exam, Spring 2025
Section: 17_A
Course Code: CIS 312
Course Title: Big Data Analytics & Applications

Time: 2 Hours **Total Marks: 40**

(The figure of the right margin indicates the marks)

(You need to answer all of the following questions)

1.	A health-tech startup is developing a smart wearable device that collects real-time data such as heart rate, sleep patterns, and physical activity. The company aims to use this data to provide personalized health insights and early detection of health issues. The data is continuously streamed from thousands of users across the country, generating large volumes of structured and unstructured data (e.g., sensor logs, audio sleep recordings, textual feedback). As a Big Data analyst, you are asked to design a high-level data strategy to manage and analyze this incoming data effectively. Based on your understanding of Big Data and its characteristics: a) Identify three major challenges that the startup might face in handling such data. b) Explain how the **5 V's** of Big Data apply to this scenario. c) Suggest two types of data analytics that could be applied to derive valuable health insights from the data, and briefly justify your choices.	3 5 2	CLO1
2.	A large dataset of 2 TB is to be processed using the Hadoop Distributed File System (HDFS). Each block in HDFS is **128 MB** in size, and the default **replication factor is 3**. a) Calculate the number of HDFS blocks that will be created to store the 2 TB file. **b)** Compute the total storage required in the Hadoop cluster to store the replicated blocks.	5 5	CLO4
3.	A large e-commerce company, ShopTrend, processes massive amounts of customer transaction data daily to generate real-time sales reports and customer behavior analytics. The company initially used Hadoop 1.0 for data processing but faced scalability issues as their data volume grew beyond 4000 nodes. To address this, they plan to upgrade to Hadoop 2.0 with YARN and MapReduce for better resource management and scalability. The company wants to implement a MapReduce job to count the frequency of product purchases across different regions, similar to the word count example discussed in the class, and leverage YARN for efficient job execution. a) Explain the limitations of Hadoop 1.0 (MR 1) that ShopTrend might have encountered, and how Hadoop 2.0 with YARN addresses these issues. b) Illustrate the YARN application workflow for executing this MapReduce job in Hadoop 2.0, highlighting the roles of the Resource Manager, Node Manager, and Application Master.	5 5	CLO2

4.	TechTrend Innovations, a rapidly growing e-commerce company, has accumulated vast amounts of data from its online sales platform, customer feedback on social media, inventory management systems, and third-party logistics providers. The company aims to leverage this data to optimize inventory, predict customer demand, and enhance marketing strategies. However, the data is stored across multiple heterogeneous systems, including relational databases, CSV files, and unstructured social media logs. The management has decided to implement a data warehouse to consolidate this data for advanced analytics and is considering whether to use an ETL or ELT process for data integration.		CLO3
	a) Explain the key differences between ETL and ELT processes in the context of TechTrend's requirements, highlighting which approach would be more suitable for their diverse data sources. Justify your recommendation regarding data compatibility, speed, and transformation needs.	5	
	b) Discuss how a Star Schema could be designed for TechTrend's data warehouse to support analysis of sales and customer behavior. Specify at least two dimensions and one fact table, including relevant attributes, and explain why this schema is appropriate for their analytics goals.	5	