UGANDA INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY
PROGRESSIVE ASSESSMENT EXAMINATIONS
ACADEMIC YEAR 2025/2026

PROGRAMME(S): DIPLOMA IN DATA SCIENCE MANAGEMENT AND ANALYTICS (DSMA)
DIPLOMA IN E-GOVERNANCE AND DIGITAL TRANSFORMATION (DEGDT)

YEAR OF STUDY: ONE SEMESTER TWO

COURSE: INTRODUCTION TO BIG DATA, CLOUD COMPUTING AND BUSINESS INTELLIGENCE

COURSE CODE : DSA1205

DATE: SUNDAY 28TH, JUNE 2026

TIME: 9:00 AM – 12:00 PM

DURATION: 3 HOURS

VENUE: LAB 1, LAB 2, LAB 3

INSTRUCTIONS:

(a) This paper consists of two sections A and B
(b) Section A carry 40 marks and is compulsory
(d) Section B is Practical: Answer any TWO questions from section B
(e) All questions in section B carry equal marks
(f) Credit will be given for use of relevant examples and illustrations

SECTION A: (40 Marks)
This section is compulsory, answer all questions

SECTION A: PART 1 (20 Marks)

1. A telecommunications company collects call records from 50 million customers daily. The traditional database system can no longer store all the data efficiently.

Which Big Data characteristic is the company experiencing?

A. Velocity
B. Volume
C. Variety
D. Veracity

2. A hospital collects patient records, X-ray images, audio recordings, and doctor's notes.

Which Big Data characteristic does this represent?

A. Volume
B. Variety
C. Velocity
D. Value

3. A retail company wants to store petabytes of sales data across hundreds of low-cost servers.

Which technology is most appropriate?

A. Microsoft Word
B. Hadoop Distributed File System (HDFS)
C. PowerPoint
D. MySQL Workbench

4. A supermarket analyzes customer purchases to identify frequently bought-together products.

Which type of analytics is being used?

A. Descriptive Analytics
B. Predictive Analytics
C. Diagnostic Analytics
D. Prescriptive Analytics

5. A bank uses historical transaction data to predict customers likely to apply for loans.

Which type of analytics is involved?

A. Diagnostic
B. Descriptive
C. Predictive
D. Prescriptive

6. A marketing company analyzes tweets, Facebook posts, and Instagram comments.

These data sources are primarily classified as:

A. Structured Data
B. Semi-Structured Data
C. Unstructured Data
D. Relational Data

7. A company needs fault-tolerant storage where data remains available even if some servers fail.

Which Hadoop feature provides this capability?

A. Replication
B. Compression
C. Normalization
D. Encryption

8. A city authority collects traffic data from road sensors every second.

What is the primary source of Big Data in this scenario?

A. Social Media
B. IoT Devices
C. Emails
D. Databases

9. A company wants to process huge datasets across multiple computers simultaneously.

Which Hadoop component performs this task?

A. HDFS
B. YARN
C. MapReduce
D. Hive

10. A logistics company tracks delivery vehicles in real time using GPS.

What type of data processing is required?

A. Batch Processing
B. Stream Processing
C. Archival Processing
D. Offline Processing

11. A company generates 10 TB of data daily, overwhelming its traditional warehouse.

Why is Big Data technology needed?

A. To reduce internet speed
B. To handle massive datasets efficiently
C. To eliminate databases
D. To remove data security controls

12. A company needs faster processing than traditional MapReduce.

Which technology should it adopt?

A. Excel
B. Spark
C. Access
D. Oracle Forms

13. A multinational company stores raw structured and unstructured data before analysis.

What storage architecture is being used?

A. Data Mart
B. Data Lake
C. Spreadsheet
D. Flat File System

14. A company analyzes customer comments to determine positive or negative opinions.

What Big Data application is this?

A. Data Mining
B. Sentiment Analysis
C. Data Replication
D. Data Compression

15. An e-commerce platform adds more servers during holiday sales to manage increased traffic.

Which Big Data feature does this demonstrate?

A. Scalability
B. Duplication
C. Compression
D. Redundancy

16. A bank develops a model that learns from past fraud cases and improves detection accuracy.

Which technology is being applied?

A. Spreadsheet Analysis
B. Machine Learning
C. Word Processing
D. Database Backup

17. A retailer discovers hidden purchasing patterns in customer transactions.

Which process is being performed?

A. Data Mining
B. Data Entry
C. Data Formatting
D. Data Backup

18. A city government collects data from CCTV cameras, traffic lights, and environmental sensors.

Which Big Data source category is represented?

A. IoT Data
B. Spreadsheet Data
C. Manual Records
D. Printed Reports

19. A healthcare provider encrypts patient data stored in Hadoop.

What objective is being achieved?

A. Data Compression
B. Data Confidentiality
C. Data Duplication
D. Data Redundancy

20. A Hadoop cluster contains one master node and several worker nodes.

What is the main advantage of this architecture?

A. Centralized Manual Processing
B. Distributed Storage and Processing
C. Reduced Data Collection
D. Elimination of Analytics

SECTION A: PART 2 (20 marks)

1. Define Big Data and explain why traditional data processing tools (like spreadsheets or standard SQL databases) are inadequate for handling it. **(04 marks)**
2. Define the concept of Distributed Computing. How does it enable the processing of massive datasets that a single machine cannot handle? **(04 marks)**
3. In the context of the 5Vs of Big Data, what does "Volume" refer to? Give a real-world example of a system or source that generates a massive volume of data. **(04 marks)**
4. Differentiate between Structured, Semi-Structured, and Unstructured data. Provide one file format/example for each category. **(04 marks)**
5. Social media platforms generate videos, images, and text posts. Classify this type of data and explain why it is challenging to analyze using traditional relational databases. **(04 marks)**

SECTION B: (60 Marks) - PRACTICAL

Each question in this section is worth **30 marks;** answer **ANY TWO** questions.

Question 1: RETAIL SALES ANALYSIS USING PIVOT TABLES (30 MARKS)
Background Scenario

Kampala Retail Hub Ltd., a fast-growing supermarket chain operating across multiple regions in Uganda, would like to analyze sales performance across regions, product categories, and salespersons. The management team requires insights to support business decision-making.

Using Microsoft Excel Pivot Tables and the dataset provided, perform the following tasks.

Task

Using the provided retail sales dataset, create Pivot Tables and answer the questions below.

(a) Total Sales by Region (6 Marks)

i. Create a Pivot Table showing the total sales for each region.

ii. Identify the region with the highest total sales and store your answer in a worksheet named: best_region

(b) Sales by Product Category (6 Marks)

i. Create a Pivot Table showing the total sales for each product category.

ii. Identify the product category with the highest sales revenue and store your answer in a worksheet named: top_category

(c) Salesperson Performance (6 Marks)

i. Create a Pivot Table showing the total sales for each salesperson, sorted from highest to lowest.

ii. Identify the top-performing salesperson and store your answer in a worksheet named: top_salesperson

(d) Sales Trend Analysis (6 Marks)

i. Create a Pivot Table showing total sales by date.

ii. Identify the date on which the highest sales occurred and store your answer in a worksheet named: best_sales_date

(e) Multi-Dimensional Analysis (6 Marks)

Create a Pivot Table showing:

Region × Product Category

Using the Pivot Table, determine:

i. Which region has the highest sales for Beverages and store the answer in: beverage_region

ii. Which region has the highest sales for Food and store the answer in: food_region

iii. Which region has the highest sales for Household Products and store the answer in: household_region

Question 2: E-COMMERCE SALES ANALYSIS USING PYTHON (30 MARKS)
Background Scenario
RYECO LTD is a fast-growing e-commerce company operating across multiple regions in Uganda. The company has provided a dataset named DSA_ecommerce_bigdata.csv containing information on sales, product categories, customer types, profit, and transaction dates.

As a Data Analyst, use Python (Pandas and Matplotlib) in a Jupyter Notebook to analyze the dataset and generate business insights.

Task
Using the dataset DSA_ecommerce_bigdata.csv, write Python code to perform the following tasks:

(a) Data Loading and Exploration (7 Marks)

i. Load the dataset into a Pandas DataFrame named df.

ii. Display the first five records of the dataset.

iii. Display the dataset structure and summary information.

(b) Data Cleaning and Feature Engineering (7 Marks)

i. Check for missing values in the dataset.

ii. Convert the Date column into DateTime format.

iii. Create a new column called Month from the Date column.

(c) Business Metrics Analysis (8 Marks)

i. Calculate the total sales and store the result in a variable called total_sales.

ii. Calculate the total profit and store the result in a variable called total_profit.

iii. Determine the region with the highest sales and store the result in best_region.

iv. Determine the product category with the highest sales and store the result in top_category.

(d) Data Visualization (8 Marks)

i. Create a bar chart showing total sales by region.

ii. Create a line chart showing monthly sales trends.

iii. Display appropriate chart titles and labels.

Question 3: BIG DATA PROCESSING USING APACHE SPARK (30 Marks)
Background Scenario

DSA_UICT has collected a large dataset named DSA_ecommerce_bigdata.csv containing information on sales transactions from different regions of Uganda. As a Big Data Analyst, you are required to use Apache Spark (PySpark) to process and analyze the dataset efficiently.

Using PySpark in a Jupyter Notebook, write code to load, process, analyze, and visualize the data.

Task

Using the dataset DSA_ecommerce_bigdata.csv, perform the following tasks:

(a) Load and Explore the Dataset (15 Marks)

i. Create a Spark Session named spark.

ii. Load the dataset into a Spark DataFrame named df.

iii. Display the schema of the dataset.

iv. Display the first five records.

(b) Data Preparation (15 Marks)

i. Check for missing values in the dataset.

ii. Convert the Date column to Date format.

iii. Create a new column called Month from the Date column.

Question 4: BIG DATA PROCESSING USING POWER BI (30 Marks)

You have been appointed as Inventory Manager at Mukwano Industries to help manage their Production inventory. Load the M_Inventory dataset into Power BI and use it to identify risks, shortages, and opportunities. Your job is to understand stock levels, identify risks, and support

restocking decisions using this data. Please add the following new columns to the existing dataset.
As part of data transformation, you are required to determine the following;

1. How much money is tied up in each product? (2 Marks)
Create a new column called "**Total Inventory value**" that calculates the total value of stock for each product by combining how many items you have and how much each item costs.

2. How much money is tied up in each product? (2 Marks)
Create a column called "**Re-Order costs"** that shows how much money is needed to restock each product up to its reorder level.

3. Which products need to be restocked? (2 Marks)
Compare the current stock quantity with the reorder level and create a column called **"Stock Status"** that tells whether the product needs restocking or not.

4. Which products are in danger of running out? (2 Marks)
Create a column called **"Critical Stock"** that flags products as critical if their stock is very low compared to the reorder level. *

5. How long has it been since we last restocked this product? (2 Marks)
Create a new column called "**Days since Last Restock**" that calculates the number of days between today and the last restock date.

6 Are we restocking regularly or not? (2 Marks)
Based on the number of days since the last restock, create a column called. "**Restock Recency**" that will group products into: Recently restocked, Moderately restocked, Not restocked for a long time.

7 Which suppliers deliver fast and which are slow?" (2 Marks)
Create a column called **"Lead Time Category"** that groups suppliers based on how many days they take to deliver (fast, normal, slow).

8. Which products have too much or too little stock? (2 Marks)
Create a column called **"Stock Level Category"** that groups products into: High stock, Medium stock, Low stock based on their quantity.

9. Are we holding too much stock? (2 Marks)
Create a column called **"Over Stock Indicator"** that identifies products where the stock is much higher than the reorder level.

10. Which products are risky to manage? (2 Marks)
Create a column called **"Risk Indicator"** that flags products as high risk if: Stock is low AND Supplier takes long to deliver.

11. Which warehouse is most important to operations? (2 Marks)
Create a column called **"Warehouse Importance"** that groups warehouses into: Main hub and Secondary locations.

DAX CALCULATIONS:

12. What is the total Stock Quantity? (2 Marks)
13. What is the total Inventory value? (2 Marks)
14. What is the average supplier lead time? (2 Marks)
15. Which warehouse has the highest stock quantity? (2 Marks)

END

EXAMINER'S NOTE TO FINALISTS:

"Exams do not measure your intelligence alone, but your discipline, preparation, and ability to perform under pressure"