Tuwaiq | AI Diploma
Course 3: Scalable Data Science
10/04/2025

Practise: Scalable Data Science with Python

Instructions: Choose the best answer (A, B, C, or D) for each question based on the provided presentation slides.

1. What is the title of Course Three in Semester One?
   A. AI Diploma
   B. Foundations of Scalable Computing
   C. Scalable Data Science with Python
   D. Distributed Computing Frameworks

2. According to the roadmap, what is the objective of Week 1: Foundations of Scalable Computing?
   A. Develop skills in GPU-accelerated data processing.
   B. Apply scalable approaches to train machine learning models.
   C. Master the fundamentals of Python data tools and understand scalable data processing principles.
   D. Configure and utilize cloud services for production deployment.

3. Which week focuses on Large-Scale Data Processing & GPU Acceleration using tools like RAPIDS, CuPy, and cuDF?
   A. Week 1
   B. Week 2
   C. Week 3
   D. Week 4

4. What is the main objective of Week 5: Cloud Infrastructure & Production Deployment?
   A. Learn distributed computing frameworks.
   B. Apply machine learning at scale using GPU acceleration.
   C. Configure cloud services and implement end-to-end data science pipelines in production.
   D. Master Python data tools fundamentals.

5. According to the Venn diagram on slide 6, Data Science sits at the intersection of which three fields?
   A. Machine Learning, Software Development, Traditional Research

B. Computer Science/IT, Math and Statistics, Domains/Business Knowledge
C. Big Data, Cloud Computing, Artificial Intelligence
D. Data Analysis, Data Mining, Data Visualization

6. Which of the following is listed as a key component of Data Science on slide 7?
A. Hardware Engineering
B. Marketing Strategy
C. Communication and Visualization Skills
D. Product Management

7. In the Data Science Lifecycle (slide 8), what step typically follows Data Preparation?
A. Business Understanding
B. Data Acquisition
C. Exploratory Data Analysis
D. Modeling

8. Which step in the Data Science Lifecycle involves assessing model performance and validating results?
A. Data Preparation
B. Modeling
C. Deployment
D. Evaluation

9. What is Data Acquisition defined as on slide 9?
A. Cleaning and transforming data.
B. The process of collecting and gathering raw data from various sources. C. Discovering patterns using automated techniques.
D. Visualizing data insights.

10. Which type of data analysis examines what might happen?
A. Descriptive
B. Diagnostic
C. Predictive
D. Prescriptive

11. What is Data Mining described as on slide 9?
A. The entire process of data analysis from start to finish.
B. A specific subset of data analysis focusing on discovering hidden patterns in large datasets.
C. The initial investigation of data using summary statistics.
D. The conversion of raw data into a more useful format.

12. What does Data Wrangling (or Data Munging) primarily address? (Slide 10) A. Collecting data from sensors.
B. Training machine learning models.

C. Cleaning, structuring, and enriching raw data; addressing inconsistencies and missing values.
D. Deploying models into production.

13. According to the comparison chart on slide 11, what kind of questions does Business Intelligence typically answer?
A. What if...? What will...? How can we...?
B. What happened...? How much did...? When did...?
C. Why did it happen...? What are the root causes...?
D. How can we make it happen...? What is the optimal action...?

14. The comparison on slide 11 suggests Data Science focuses more on the ____ compared to Business Intelligence.
A. Past
B. Present
C. Future
D. Historical Data Only

15. Slide 12 states that data scientists are NOT domain experts, but they must...? A. Know everything about the domain.
B. Work together with domain experts.
C. Ignore domain knowledge.
D. Become domain experts quickly.

16. The "Scale Challenge" slides (13-17) show examples of large data generation from sources EXCEPT:
A. DNA sequencing labs
B. Radio telescope arrays
C. Social network interactions (Twitter)
D. Personal diaries

17. Slide 18 mentions that global data creation is projected to exceed how many zettabytes?
A. 18 Zettabytes
B. 180 Zettabytes
C. 1.8 Zettabytes
D. 1800 Zettabytes

18. Which scaling dimension involves increasing resources (RAM, CPU, GPU) on a single machine? (Slide 19)
A. Horizontal Scaling (Scale Out)
B. Vertical Scaling (Scale Up)
C. Algorithm Efficiency
D. Data Architecture

19. Distributing processing across multiple machines is known as? (Slide

19) A. Vertical Scaling (Scale Up)
B. Horizontal Scaling (Scale Out)
C. Algorithm Efficiency
D. Data Architecture Optimization

20. What does the CPU (Central Processing Unit) primarily handle, according to slide 20?
A. Many simple calculations all at once.
B. General calculations and tasks, good at a few complex tasks sequentially. C. Only graphics and video rendering.
D. Storing data actively being used.

21. What is the GPU (Graphics Processing Unit) excellent at doing? (Slide 20) A. Running the operating system efficiently.
B. Handling complex, sequential decision-making.
C. Doing many simple calculations all at once (parallel processing). D. Long-term data storage.

22. What is the role of RAM (Random Access Memory) described as on slide 21? A. The "brain" of the computer.
B. Long-term storage for files.
C. The "workspace" or "short-term memory" holding actively used data.
D. Specialized unit for graphics.

23. What happens when a computer runs out of RAM? (Slide 21) A. It speeds up processing.
B. It shuts down immediately.
C. Everything slows down as it uses slower storage (like a hard drive) as a substitute.
D. It automatically orders more RAM.

24. Which hardware upgrade is an example of Vertical Scaling? (Slide 22) A. Setting up a Hadoop cluster.
B. Upgrading RAM from 16GB to 64GB.
C. Deploying Spark across multiple machines.
D. Using AWS EMR.

25. Using CuPy instead of NumPy for array operations is an example of what kind of optimization for Vertical Scaling? (Slide 23)
A. Multi-threading
B. Memory-Efficient Algorithms
C. GPU-Accelerated Libraries
D. RAM Expansion

26. What is a key limitation of Vertical Scaling? (Slide 24)
A. It requires complex network configuration.
B. It eventually hits physical and economic limits.

C. It is only suitable for small datasets.
D. It cannot utilize GPUs.

27. Which technology is associated with Horizontal Scaling by setting up a network of commodity servers running HDFS and MapReduce? (Slide 25) A. Dask Clusters
B. Spark Clusters
C. Hadoop Clusters
D. Vertical Scaling

28. Which cloud-based solution is mentioned for Horizontal Scaling that allows dynamically scaling clusters based on workload? (Slide 26)
A. Google Dataproc/BigQuery
B. AWS EMR (Elastic MapReduce)
C. Local Spark installation
D. Sharding

29. What database scaling technique involves splitting databases across multiple servers by partitioning keys? (Slide 27)
A. Replication
B. Sharding
C. Vertical Scaling
D. Caching

30. Distributing large machine learning models across multiple GPUs/machines where parameters don't fit in a single GPU's memory is called? (Slide 28) A. Data-Parallel Processing
B. Model-Parallel Training
C. Vertical Scaling
D. Sharding

31. Slide 29 illustrates training Large Language Models (LLMs) requires significant resources, mentioning approximately how many GPUs?
A. 6 GPUs
B. 60 GPUs
C. 600 GPUs
D. 6,000 GPUs

32. Why has Python become the de facto language for data science, according to slide 30?
A. It's the oldest programming language.
B. It has readability, an extensive ecosystem, community support, and versatility.
C. It's exclusively designed for GPU computing.
D. It requires minimal hardware resources.

33. Which Python library is primarily used for numerical computing with

multi-dimensional arrays? (Slide 31)
A. Pandas
B. Matplotlib
C. Scikit-learn
D. NumPy

34. Which Python library is mentioned for data manipulation and analysis? (Slide 31) A. NumPy
B. Pandas
C. Seaborn
D. SciPy

35. Dask is described as which type of library for scaling Python? (Slide 32) A. GPU-accelerated data science library
B. Parallel computing library
C. JIT compiler for numeric functions
D. Data visualization tool

36. What does import numpy as np achieve in Python? (Slide 33) A. It defines a new function called numpy.
B. It installs the NumPy library.
C. It imports the NumPy library and assigns it the alias 'np'.
D. It runs a NumPy calculation.

37. How do you create a NumPy array of zeros with a specific shape and data type? (Slide 35)
A. np.create_zeros(shape, type)
B. np.zeros(shape, dtype=...)
C. np.array(0, shape)
D. numpy.zeros(shape)

38. What is the primary use of the Pandas library mentioned on slide 36? A. Numerical computations on arrays.
B. Plotting and visualization.
C. Importing and managing datasets (Data Analysis, Manipulation, Visualization). D. Machine learning algorithms.

39. In Pandas, what is a 1D labeled (indexed) array called? (Slide 38) A. DataFrame
B. Series
C. Matrix
D. Vector

40. What is a 2D labeled (indexed) matrix in Pandas called? (Slide 38) A. Series
B. DataFrame

C. Array
D. List

41. How would you select only column 'x1' from a Pandas DataFrame named df? (Slide 39)
A. df.select('x1')
B. df['x1']
C. df.column('x1')
D. df.get('x1')

42. Which library is commonly used for plotting and visualizing data in Python, as shown on slide 40?
A. NumPy
B. Pandas
C. Matplotlib
D. Scikit-learn

43. What function in Matplotlib's pyplot (aliased as plt) is used to create a basic 2D line graph? (Slide 41)
A. plt.bar()
B. plt.hist()
C. plt.scatter()
D. plt.plot()

44. How do you display a Matplotlib plot after defining it? (Slide 41) A. plt.render()
B. plt.display()
C. plt.show()
D. plt.execute()

45. What function adds a legend to a Matplotlib plot, typically used when plotting multiple lines? (Slide 43)
A. plt.labels()
B. plt.title()
C. plt.legend()
D. plt.annotate()

46. Compared to CPUs, GPUs typically have: (Slides 47-49)
A. Fewer, more powerful cores.
B. Larger cache per core.
C. Many simpler cores.
D. Optimization for sequential processing.

47. GPUs are primarily designed for what type of operations, making them suitable for data science? (Slide 48)
A. Low-latency operations

B. Complex branch prediction
C. Sequential task execution
D. High-throughput, parallel, mathematical operations

48. CPUs excel at tasks requiring: (Slide 50)
A. Highly parallel computations.
B. Complex decision-making and sequential processes.
C. Working with extremely large datasets that don't fit in cache.
D. SIMD operations.

49. What technology introduced by NVIDIA enabled General-Purpose computing on GPUs (GPGPU)? (Slide 51)
A. Programmable Shaders
B. Tensor Cores
C. CUDA
D. cuDNN

50. According to slide 52, GPU computing performance growth is significantly _____ than CPU performance growth.
A. Slower
B. Faster
C. Equal to
D. More erratic than

51. In modern GPU architecture, what are the basic processing blocks called? (Slide 53)
A. CUDA Cores
B. Tensor Cores
C. Streaming Multiprocessors (SMs)
D. Global Memory (VRAM)

52. What is the purpose of Shared Memory in the GPU Memory Hierarchy? (Slide 53)
A. Main large storage accessible by all components.
B. Ultra-fast storage for individual thread variables.
C. Fast memory shared between threads in the same block for communication. D. Automatic buffer storage to reduce memory access times.

53. In the CUDA programming model, what is a 'Block'? (Slide 54) A. An individual execution unit.
B. A collection of blocks that execute the same kernel.
C. The main CPU controlling the GPU.
D. A group of threads that can communicate via shared memory.

54. Which of the following is listed as a reason to use GPU acceleration in Data Science? (Slide 55)
A. Increased latency

B. Lower energy efficiency
C. Speed (10-100x faster for suitable algorithms)
D. Inability to process large datasets

55. Which common Machine Learning algorithm type is listed as being accelerated by GPUs? (Slide 56)
A. Rule-based systems
B. Decision Stumps only
C. Neural Networks and Deep Learning
D. Manual data entry

56. Which library is part of the Python GPU Ecosystem specifically for array computation? (Slide 57)
A. RAPIDS (cuDF, cuML)
B. Dask-CUDA
C. CuPy, PyTorch, TensorFlow
D. CUDA Python

57. According to slide 58, which type of NVIDIA GPU (like Tesla/A100/H100) is typically found in High-Performance (Server/Cloud) environments? A. Entry-level (GTX/RTX Consumer Cards)
B. Professional (RTX A-series/Quadro)
C. Integrated laptop GPUs
D. High-Performance (Tesla/A100/H100)

58. What is "Big Data" generally defined as? (Slide 62)
A. Any data stored digitally.
B. Data sets so large or complex that traditional processing applications are insufficient.
C. Data measured only in Petabytes or larger.
D. Data used exclusively for machine learning.

59. The "5 Vs of Big Data" include Volume, Velocity, Variety, Veracity, and...? (Slide 63)
A. Validity
B. Value
C. Visibility
D. Volatility

60. Which 'V' of Big Data refers to the trustworthiness and quality of data? (Slide 63) A. Volume
B. Velocity
C. Variety
D. Veracity

61. What type of data includes JSON and XML? (Slide 66)
A. Structured Data

B. Unstructured Data
C. Semi-structured Data
D. Relational Data

62. What percentage of all data is estimated to be Unstructured Data (like text, audio, video)? (Slide 66)
A. 20%
B. 50%
C. 80%
D. 100%

63. What common data quality issue is mentioned under "Veracity"? (Slide 68) A. Perfect formatting
B. High speed generation
C. Missing values, duplicate records, inconsistent formats
D. Small data volume

64. The shift in handling Big Data involves moving from scaling up to...? (Slide 70) A. Scaling down
B. Scaling out
C. Scaling in
D. Scaling vertically

65. What does the CAP theorem state about distributed systems? (Slide 72) A. You can always achieve Consistency, Availability, and Partition Tolerance simultaneously.
B. Consistency is the most important property.
C. You can only guarantee 2 out of the 3 properties (Consistency, Availability, Partition Tolerance).
D. Partition Tolerance is optional for modern systems.

66. In distributed systems requiring Partition Tolerance, what trade-off must typically be made according to the CAP theorem? (Slide 72)
A. Speed vs. Cost
B. Storage vs. Compute
C. Consistency vs. Availability
D. Durability vs. Performance

67. Which distributed processing programming model involves Map, Shuffle, and Reduce phases? (Slide 73 & 75)
A. Bulk Synchronous Parallel (BSP)
B. Directed Acyclic Graphs (DAGs)
C. MapReduce
D. Stream Processing

68. Stream processing primarily deals with: (Slide 74)
A. Large batches of historical data processed at scheduled intervals.

B. Data processed in real-time or near real-time as it arrives.
C. Small, static datasets.
D. Data that does not require transformation.

69. What is the primary advantage of the MapReduce paradigm listed on slide 75? A. Real-time processing capability.
B. Suitability for small datasets only.
C. Simple programming model, automatic parallelization, fault tolerance. D. Requires manual management of data distribution.

70. What is Distributed Computing defined as? (Slide 78/80)
A. Running computations on a single powerful machine.
B. Using multiple GPUs on one computer.
C. A paradigm where components on networked computers communicate and coordinate to achieve a common goal.
D. Storing data in a centralized database.

71. What is a "Cluster" in the context of distributed systems? (Slide 83) A. A single computer/server in the network.
B. A collection of nodes working together.
C. A redundant copy of data.
D. A strategy for dividing data.

72. Which distributed architecture pattern involves a central coordinator assigning tasks to workers? (Slide 84)
A. Peer-to-Peer
B. Master-Worker (Master-Slave)
C. Hybrid Approaches
D. Client-Server (non-distributed context)

73. What is a major disadvantage of the Master-Worker architecture? (Slide 84) A. Complex coordination
B. No single point of failure
C. Single point of failure at the master
D. High resilience

74. What is the "80/20 Rule of Data Science" mentioned on slide 86? A. 80% analysis, 20% data preparation.
B. 80% modeling, 20% deployment.
C. 80% data preparation, 20% actual analysis and modeling.
D. 80% deployment, 20% monitoring.

75. Which is NOT listed as a reason why preprocessing at scale is different? (Slide 86)
A. Can't load all data into memory.
B. Performance bottlenecks are amplified.

C. Can manually inspect all records easily.
D. New failure modes emerge.

76. What is Apache Spark defined as? (Slide 94)
A. A Python library for plotting.
B. An open-source, distributed computing system for fast, large-scale data processing.
C. A relational database management system.
D. A GPU-accelerated machine learning library.

77. What is a key feature of Spark that makes it faster than Hadoop MapReduce for certain operations? (Slide 94/95)
A. Disk-based computation only
B. In-memory computation
C. Single-machine processing
D. Lack of fault tolerance

78. What is PySpark? (Slide 98)
A. The core execution engine of Spark.
B. A machine learning library for Spark.
C. The Python API for Apache Spark.
D. A structured data processing module in Spark.

79. What is the fundamental data structure in Spark described as an immutable, distributed collection of objects? (Slide 101 - pg 1 of unit 2 pdf)
A. DataFrame
B. Dataset
C. Resilient Distributed Dataset (RDD)
D. SparkContext

80. RDD operations are categorized into which two types? (Slide 102 - pg 2 of unit 2 pdf)
A. Read and Write
B. Map and Reduce
C. Transformations and Actions
D. Compile and Execute

81. Which type of RDD operation is lazily evaluated? (Slide 102 - pg 2 of unit 2 pdf) A. Actions
B. Transformations
C. Collect
D. Count

82. What triggers the execution of RDD transformations? (Slide 102 - pg 2 of unit 2 pdf)
A. Creating the RDD
B. Calling another transformation

C. Calling an Action
D. Saving the code

83. What is a Spark DataFrame? (Slide 104 - pg 4 of unit 2 pdf)
A. An unstructured collection of objects.
B. A distributed collection of data organized into named columns, similar to a table.
C. The main entry point for Spark functionality.
D. A graph computation engine.

84. Which Spark component is specifically designed for machine learning? (Slide 107 - pg 7 of unit 2 pdf)
A. Spark SQL
B. Structured Streaming
C. GraphX
D. MLlib

85. What is Dask? (Slide 114 - pg 14 of unit 2 pdf)
A. A distributed SQL database.
B. A Python-native parallel computing library.
C. An alternative API for Apache Spark.
D. A GPU-only machine learning framework.

86. Which Dask component provides parallel versions of NumPy arrays? (Slide 116 - pg 16 of unit 2 pdf)
A. dask.dataframe
B. dask.bag
C. dask.array
D. dask.delayed

87. Dask operations are typically _____, meaning computation happens only when needed (e.g., calling .compute()). (Slide 116, 118 - pg 16, 18 of unit 2 pdf) A. Eager
B. Lazy
C. Immediate
D. Synchronous

88. What is dask.delayed used for? (Slide 116, 121 - pg 16, 21 of unit 2 pdf) A. Creating parallel NumPy arrays.
B. Creating parallel Pandas DataFrames.
C. Wrapping custom Python functions/code for lazy execution and parallelization.
D. Executing tasks immediately.

89. Compared to Apache Spark, Dask generally has: (Slide 129 - pg 29 of unit 2 pdf) A. Higher startup time due to JVM overhead.
B. Lower overhead and is more lightweight.

C. Centralized DAG scheduler only.
D. Less integration with the Python ecosystem.

90. Which framework, Spark, Dask, or Ray, is specifically highlighted for native capabilities like hyperparameter tuning (Tune) and reinforcement learning (RLlib)? (Slide 135 - pg 35 of unit 2 pdf)
A. Apache Spark
B. Dask
C. Ray
D. All three equally

91. What is CuPy designed as an alternative to? (Slide 145 - pg 45 of unit 2 pdf) A. Pandas
B. Scikit-learn
C. NumPy
D. Spark
92. What hardware does CuPy leverage for acceleration? (Slide 146 - pg 46 of unit 2 pdf)
A. CPUs
B. FPGAs
C. GPUs (via CUDA)
D. TPUs

93. What is a key advantage of Lazy Evaluation? (Slide 154 - pg 54 of unit 2 pdf) A. Immediate results for interactive exploration.
B. Simpler debugging of complex transformations.
C. Optimization opportunities (reordering, eliminating operations) and memory efficiency.
D. Always faster for simple, one-off calculations.

94. What is an Execution Plan in the context of lazy evaluation systems like Spark or Dask? (Slide 156 - pg 56 of unit 2 pdf)
A. The source code written by the user.
B. A blueprint describing operations, order, optimization, and resource allocation.
C. The final result of the computation.
D. The hardware configuration.

95. What is RAPIDS? (Slide 194 - pg 94 of unit 3 pdf)
A. A type of CPU.
B. A distributed database.
C. NVIDIA's suite of open-source libraries for end-to-end data science pipelines on GPUs.
D. A cloud computing platform.

96. Which core component of RAPIDS provides a GPU DataFrame library with

a Pandas-like API? (Slide 194, 196 - pg 94, 96 of unit 3 pdf)
A. cuML
B. cuGraph
C. cuDF
D. cuSignal

97. Compared to Pandas, cuDF is generally faster for: (Slide 198 - pg 98 of unit 3 pdf)
A. Small datasets where CPU overhead matters.
B. Large datasets due to GPU acceleration.
C. Operations that cannot be parallelized.
D. Environments without GPUs.

98. What is cuML? (Slide 223 - pg 123 of unit 4 pdf)
A. A GPU-accelerated graph analytics library.
B. A GPU-accelerated DataFrame library.
C. An open-source GPU-accelerated machine learning library, part of RAPIDS. D. A CPU-based machine learning library.

99. cuML provides a similar API to which popular CPU-based machine learning library, easing transition? (Slide 223 - pg 123 of unit 4 pdf)
A. TensorFlow
B. PyTorch
C. Scikit-learn
D. Keras

100. What is a primary limitation of using cuML? (Slide 232 - pg 132 of unit 4 pdf)
A. It only works on CPUs.
B. It requires an NVIDIA GPU.
C. It is slower than Scikit-learn for large datasets.
D. It has a completely different API from Scikit-learn.

- **Name: Muath Alsaeed**
- **Grade:**
- **Date:**