NVIDIA-Certified Professional: Accelerated Data Science 认证考试学习指南

NVIDIA-Certified Professional: Accelerated Data Science 认证考试学习指南

目录

针对 NVIDIA-Certified Professional：Accelerated Data Science (NCP-ADS) 认证考试，本学习指南包含每个认证主题的介绍，并推荐相关培训课程和阅读资料。

查看所有 NVIDIA 认证，请单击此处。

考试适用人群

加速数据科学专业人员，利用前沿 GPU 加速技术来优化数据科学工作流 — 从提取、转换和加载 (ETL) 流程到机器学习模型部署。考生需具备专业的数据科学、机器学习和 GPU 计算背景，并热衷于突破数据分析和建模的极限。考生需能够实施高性能数据解决方案，利用先进工具和技术显著提高数据处理、分析和模型训练工作流的速度和效率。

相关工作职责示例

1. 加速数据科学：使用先进工具和技术来加速数据科学流程，专注于通过 GPU 加速来优化性能
2. GPU 加速：利用 GPU 计算大幅缩短数据处理、分析和模型训练所需的时间
3. 提取、转换、加载 (ETL)：高效管理和转换大型数据集，借助加速 ETL 流程进行分析
4. 数据分析：进行全面的数据分析，利用加速工作流管理大规模数据，提取有意义的洞察
5. 特征工程：创建和优化特征以提高模型性能，利用 GPU 加速方法简化流程
6. 数据建模：开发并调优数据模型，采用先进的 GPU 加速算法提高精度和速度
7. 图形分析：使用 GPU 加速工具创建和分析图形数据。实施和优化基于图形的数据结构和算法，用于建立数据关系并执行网络分析
8. 数据预处理：使用加速技术执行数据清理、转换和标准化，准备数据进行分析和建模
9. MLOps：将机器学习模型集成到生产环境中，通过加速工作流确保可扩展性和效率
10. GPU 加速机器学习：使用 GPU 加速库设计、训练和评估机器学习模型，以缩短训练时间并提高性能
11. 数据可视化：创建高质量可视化效果，利用加速工具处理大型数据集，有效传达数据驱动的洞察
12. 时序分析：应用先进的时间序列预测方法，利用加速技术更快、更准确地进行预测
13. GPU 加速数据操作：使用 GPU 加速库高效处理大型数据集，确保在数据处理任务中实现最佳性能。利用 cuDF 高效完成数据操作、特征提取和转换任务
14. 数据清理：执行彻底的数据清理，确保高质量数据输入，使用加速方法管理大型复杂数据集

建议具备的知识和经验

1. 两到三年的加速数据科学实践经验
2. 扎实的的机器学习和 GPU 加速计算基础
3. 具有基于 GPU 的优化策略和加速数据操作技术方面的经验
4. 深入了解端到端数据科学工作流 — 从数据准备和清理到模型开发和部署，利用 GPU 加速来提高性能和效率

认证主题和参考资料

数据分析：考试权重 14%

执行时序分析检测异常情况，进行趋势可视化展示，并执行探索性数据分析 (EDA) 以获取见解。使用 cuGraph 等工具来分析图形数据，并判断数据符合大数据的条件，应用适当的加速方法实现高效处理。这些步骤有助于理解和优化数据，以便做出更合理的决策。

1.1 检测时序数据集中的异常情况
1.2 进行时序分析
1.3 使用 cuGraph 等工具创建和分析图形数据
1.4 确定大数据的数据量，或加速的时间和方式
1.5 执行探索性数据分析 (EDA)
1.6 时序数据可视化

培训推荐（可选）

- 在线自主培训《加速端到端数据科学工作流》(查看课程)
- 讲师指导的培训班《加速端到端数据科学工作流》(查看课程大纲)
- 讲师指导的培训班《通过高效工作流提升数据科学成果》(查看课程大纲)

阅读内容推荐

- Supercharge Graph Analytics at Scale With GPU-CPU Fusion for 100x Performance
- Exploratory Data Analysis in Python
- Anomaly Detection for Time-Series Data: Anomaly Types
- NYC Taxi Data Using dask_cudf
- Beginner's Guide to GPU-Accelerated Graph Analytics in Python
- NetworkX Tutorial
- Big Data: A Revolution That Will Transform How We Live, Work, and Think by Viktor Mayer-Schönberger and Kenneth Cukier
- Anomaly Detection: A Survey by Varun Chandola, Arindam Banerjee, and Vipin Kumar
- Forecasting: Principles and Practice by Rob J. Hyndman and George Athanasopoulos
- Iglewicz, B., & Hoaglin, D. C. (1993). How to Detect and Handle Outliers. Sage Publications.
- Zimek, A., & Schubert, E. (2017). A Survey on Evaluation Methods for Anomaly Detection. ACM Computing Surveys, 50(3), 1-34.

数据操作和软件专业知识：考试权重 19%

利用加速处理流程、数据缓存和分布式数据处理框架设计并实施高效 ETL 工作流，用于管理大数据。利用 Dask 实施数据并行，实现多 GPU 扩展，同时借助 DLProf 等工具分析深度学习模型，并优化性能。考生需能够为不同规模的数据集选择最佳数据处理库，以及使用 cuDF 执行 GPU 加速数据转换和标准化。有效的 GPU 内存管理以及在单节点和多节点系统中对大规模数据集进行推理扩展也是关键能力。

2.1 使用加速 ETL 流程设计和实施 ETL 工作流
2.2 实施数据缓存，减少混洗
2.3 使用分布式数据处理框架来处理大数据
2.4 使用 Dask 实施数据并行，实现多 GPU 扩展
2.5 使用 DLProf 等工具分析深度学习模型
2.6 确定针对不同数据集规模使用的最佳数据处理库
2.7 使用 cuDF 进行数据转换和标准化
2.8 有效管理和分配 GPU 显存以提高性能，利用策略优化内存
2.9 在单节点和多节点系统中扩展大规模数据集推理

培训推荐（可选）

> 在线自主培训《加速端到端数据科学工作流》(查看课程)
> 讲师指导的培训班《加速端到端数据科学工作流》(查看课程大纲)
> 讲师指导的培训班《通过高效工作流提升数据科学成果》(查看课程大纲)

阅读内容推荐

> Accelerating ETL on KubeFlow With RAPIDS™
> FAQ and Known Issues
> Dask cuDF Best Practices
> NVIDIA Triton™ Inference Server
> Categorical Features in XGBoost Without Manual Encoding
> User Guide
> Unlocking Multi-GPU Model Training with Dask XGBoost
> RAPIDS on Databricks: A Guide to GPU-Accelerated Data Processing
> ETL Principles
> Accelerating Apache Spark 3.0 With GPUs and RAPIDS

数据准备：考试权重 17%

使用 cuDF 和 pandas 清理和预处理数据，对数据进行转换和标准化，确保特征一致性。使用 cuDF/RAPIDS 生成合成数据增强数据集，同时识别并获取相关数据集。此外，考生还需能够监控数据处理工作流，检测瓶颈，确保高效组织、处理和存储数据集，以便进行后续分析。

3.1 使用 cuDF 和 pandas 执行数据清理和预处理
3.2 进行数据转换和标准化
3.3 根据需要进行数据标准化，确保特征一致性
3.4 识别并获取数据集
3.5 监控数据处理工作流，识别数据处理瓶颈
3.6 处理、组织和存储数据集

培训推荐（可选）

> 在线自主培训《加速端到端数据科学工作流》(查看课程)
> 讲师指导的培训班《加速端到端数据科学工作流》(查看课程大纲)
> 讲师指导的培训班《通过高效工作流提升数据科学成果》(查看课程大纲)

阅读内容推荐

> Synthetic Data for Deep Learning
> NVIDIA NeMo™ Curator for Developers
> Optimizing Access to Parquet Data With fsspec
> Curating Non-English Datasets for LLM Training With NVIDIA NeMo Curator
> Working With JSON Data
> Faster Resampling With Imbalanced-Learn and cuML
> Machine Learning Frameworks Interoperability, Part 2: Data Loading and Data Transfer Bottlenecks
> Data Science: Understanding Feature Scaling in Machine Learning
> Encoding and Compression Guide for Parquet String Data Using RAPIDS
> Preprocessing Data
> The Importance of Feature Scaling
> Dealing With Outliers Using Three Robust Linear Regression Models
> Data Loading
> Best Practices for Monitoring Data Pipeline Performance

GPU 和云计算：考试权重 16%

利用 GPU 加速优化数据科学工作流，包括使用 cuGraph 分析图形数据，并提高各流程性能。遵循 CRISP-DM 方法论，使用 Docker 或 Conda 等框架管理软件依赖，以及为特征选择最佳数据类型。考生还需能够进行基准测试，比较各种框架的性能，了解何时以及如何使用 AWS、GCP 或 Databricks 等平台将数据处理工作流扩展到云端。

4.1 使用 cuGraph 等 GPU 加速工具来分析图形数据
4.2 通过 GPU 加速优化数据科学流程性能
4.3 描述、遵循并执行 CRISP-DM 流程
4.4 利用依赖管理框架 (如 Docker 和 Conda) 来解决软件版本冲突
4.5 确定每个特征的最优数据类型选择
4.6 通过设计和实施基准测试来比较各框架的性能

培训推荐（可选）

> 在线自主培训《加速端到端数据科学工作流》(查看课程)
> 讲师指导的培训班《加速端到端数据科学工作流》(查看课程大纲)
> 讲师指导的培训班《通过高效工作流提升数据科学成果》(查看课程大纲)

阅读内容推荐

> NetworkX Introduces Zero-Code-Change Acceleration Using NVIDIA cuGraph
> Managing Environments
> What Is Docker?
> Installing Conda
> v1.1 Results—Inference: Data Center
> LLM Inference Performance Engineering: Best Practices
> cuGraph Introduction
> Dealing With Small Files Issues on S3: A Guide to Compaction
> Use Cases of Docker: Rollback and Version Control
> RAPIDS 24.04 Release
> Remote Direct-Memory Access (RDMA)
> Data Tiering
> Autoscaling Groups of Instances
> CPU vs. GPU for Machine Learning

机器学习：考试权重 15%

通过高效分割数据并进行标准化来执行特征工程，确保特征一致性。在 GPU 加速机器学习领域，任务包括确定何时对大数据应用加速技术，快速进行实验以兼顾模型精度与性能，以及优化超参数。在单 GPU 和多 GPU 场景中，训练机器学习模型至关重要，同时应使用批处理和混合精度等技术来优化 GPU 内存。此外，深度学习模型训练涉及使用 DLProf 等工具对模型进行分析，提高性能和效率。

5.1 分割数据，选择最佳分割类型和方法
5.2 根据需要进行数据标准化，确保特征的一致性
5.3 识别数据量何时达到大数据标准，或确定加速的时间和方式
5.4 执行快速实验，兼顾模型精度与推理性能
5.5 优化机器学习模型的超参数
5.6 针对单 GPU 和多 GPU 场景训练机器学习模型
5.7 使用 GPU 显存优化技术 (例如批处理和混合精度) 来训练机器学习模型
5.8 使用 DLProf 等工具来分析深度学习模型

培训推荐（可选）

> 在线自主培训《加速端到端数据科学工作流》(查看课程)
> 讲师指导的培训班《加速端到端数据科学工作流》(查看课程大纲)
> 讲师指导的培训班《通过高效工作流提升数据科学成果》(查看课程大纲)

阅读内容推荐

> What Are the Benefits and Drawbacks of Batch Processing in Machine Learning?
> Performance Optimization
> Model Parallelism
> Parallel Training Methods for AI Models: Unlocking Efficiency and Performance
> Stratified Sampling: You May Have Been Splitting Your Dataset All Wrong
> Considerations for Hyperparameter Tuning
> Feature Scaling in Machine Learning
> Overfitting vs. Underfitting in Machine Learning
> AI Performance Tuning: Answer, Benchmark, Test
> Parallelism
> PyTorch Multi-GPU: 4 Techniques Explained
> Profiling and Optimizing Deep Neural Networks With DLProf and PyProf
> Feature Scaling
> Mixed-Precision Training
> Methods and Tools for Efficient Training on a Single GPU
> Hyperparameter Tuning Guide
> The Importance of Feature Scaling
> GPU Data Analytics: Accelerating Big Data Processing
> AI Inference Performance

MLOps：考试权重 19%

在 MLOps 中执行、优化和部署机器学习模型，包括确定最佳数据类型并评估内存需求。为工作负载选择最佳加速器，实施并行处理，并通过基准测试来优化 GPU 加速工作流。为推理工作流制定数据加载规划，在 Triton 上部署模型，并通过加密确保数据安全和合规。实施持续集成和部署 (CI/CD) 工作流以简化流程，通过缓存来缩短数据加载时间并优化机器学习工作流。

6.1 确定每个特征的最佳数据类型选择
6.2 评估和验证数据集的内存占用大小
6.3 比较所需内存与设备可用内存
6.4 执行基准测试并优化不同 GPU 加速工作流
6.5 在生产环境中部署模型并进行监控

培训推荐（可选）

> 在线自主培训《加速端到端数据科学工作流》(查看课程)
> 讲师指导的培训班《加速端到端数据科学工作流》(查看课程大纲)
> 讲师指导的培训班《通过高效工作流提升数据科学成果》(查看课程大纲)

阅读内容推荐

> Machine Learning Model Monitoring: Best Practices
> Deep Learning Specialization
> Triton Response Cache
> Protecting Sensitive Data and AI Models With Confidential Computing
> Improving GPU Memory Oversubscription Performance
> NVIDIA Triton Inference Server
> Data Types
> 10 Minutes to cuDF
> Using Multiple GPUs and Multiple Nodes
> Lustre File System
> Parallelism
> Mastering LLM Techniques: Inference Optimization
> NVIDIA TensorRT™
> vLLM
> A Guide to Quantization in LLMs
> Batch Sizes
> GPU Memory Essentials for AI Performance
> Even Faster and More Scalable UMAP on the GPU With RAPIDS cuML
> Estimator Intro
> Boosting Data Ingest Throughput With NVIDIA GPUDirect® Storage and RAPIDS cuDF
> Optimization
> Deploying Machine Learning Models on NVIDIA Triton Inference Server

遇到问题?

问题咨询，请发邮件至 dlichina@nvidia.com