NVIDIA-Certified Associate: Accelerated Data Science

NVIDIA-Certified Associate: Accelerated Data Science

Contents

Job Description

The accelerated data science associate is an early-career practitioner who uses GPU-accelerated tools to prepare data, build machine learning models, and construct reproducible pipelines on modern data platforms. The role focuses on end-to-end workflows that move from raw data to evaluated models ready for deployment.

Typical responsibilities include integrating and cleaning data, building GPU-accelerated ETL workflows, and engineering features for numerical and categorical variables. The practitioner must understand how to handle class imbalance, apply dimensionality reduction and sampling strategies, and store data efficiently in formats such as Parquet.

On the modeling side, the associate uses RAPIDS™ libraries like cuML and XGBoost to train regression, classification, and clustering models, evaluate performance with appropriate metrics, and apply cross-validation and hyperparameter tuning. They design and automate data science pipelines, mitigate underfitting and overfitting, and build reproducible workflows that scale across hardware.

The role also requires foundational knowledge of Python for data analysis, GPU acceleration concepts, CPU versus GPU workloads, and the end-to-end data science lifecycle. Introductory MLOps skills such as monitoring pipelines, tracking experiments, managing artifacts, and benchmarking workflows are essential, along with basic software and environment management (Conda, pip, Docker, Git) to ensure reproducibility in multi-user environments.

Key Responsibilities

1. Perform data integration, cleaning, feature engineering, and quality handling using cuDF, pandas, and GPU-accelerated ETL tools..
2. Train and evaluate GPU-accelerated machine learning models (regression, classification, clustering) using cuML and XGBoost.
3. Design, automate, and document end-to-end data science pipelines with RAPIDS and Dask.
4. Conduct exploratory data analysis, descriptive statistics, visualization, and hypothesis testing to derive insights.
5. Understand foundational data science concepts and be able to accelerate workloads when appropriate.
6. Implement introductory MLOps practices for monitoring, experiment tracking, model management, and hardware benchmarking.
7. Work with advanced structures such as time series and graphs using GPU-accelerated tools.
8. Manage software environments, dependencies, and version control to support reproducible GPU data science projects.

Recommended Qualifications and Experience

1. Familiarity with Python for data analysis, including NumPy, pandas, and Jupyter notebooks
2. Exposure to GPU-accelerated tools and frameworks such as RAPIDS, cuDF, cuML, and GPU-enabled XGBoost
3. Hands-on experience with at least one workflow that includes data ingestion, cleaning, feature engineering, model training, and evaluation
4. Basic experience with experiment tracking, configuration management, or model versioning tools
5. Comfort working with Linux environments, environment managers (Conda/pip), and basic Git workflows.

Certification Topics and References

Data Manipulation and Preparation – 23%

- Data integration, joining, and manipulation using cuDF and pandas
- Data cleaning, quality handling, and governance compliance
- GPU-accelerated ETL workflows with RAPIDS, Dask, or Spark
- Feature engineering for numerical and categorical variables
- Handling class imbalance and generating synthetic data
- Dimensionality reduction and data sampling
- Efficient processing and storage with Parquet and modern frameworks

Recommended NVIDIA Course and Suggested Readings

NVIDIA Course

- Accelerating End-to-End Data Science Workflows
- Best Practices in Feature Engineering for Tabular Data With GPU Acceleration

Suggested Readings

- NVIDIA cuDF: GPU-Accelerated DataFrames
- GPU-Accelerated DataFrames for pandas Users
- RAPIDS cuDF for Feature Engineering
- Best Practices for Multi-GPU Data Analysis Using RAPIDS With Dask
- Techniques to Deal With Class Imbalance

Machine Learning with RAPIDS – 16%

- GPU-accelerated model training with cuML and XGBoost
- Regression, classification, and clustering techniques
- Model evaluation, comparison, and generalization assessment
- Hyperparameter tuning and optimization
- Cross-validation methods
- Performance metrics and confusion matrix interpretation

Recommended NVIDIA Course and Suggested Readings

NVIDIA Course

- Accelerating End-to-End Data Science Workflows

Suggested Readings

- Accelerated Data Analytics: ML With GPU-Accelerated pandas and scikit-learn
- NVIDIA cuML Brings Zero-Code-Change Acceleration to scikit-learn
- XGBoost/RandomForest GPU HPO With Dask
- Scikit-learn: Common Pitfalls
- Confusion Matrix

Data Science Pipelines and Workflow Automation – 13%

- End-to-end data science pipeline design
- Feature engineering, selection, and transformation for model improvement
- Mitigating underfitting and overfitting through model and feature adjustments
- Dataset augmentation and integration for enhanced training data
- Automation and scalability of data science workflows
- Building reproducible pipelines with RAPIDS and Dask
- Enable a data flywheel where production data continuously improves deployed models

Recommended NVIDIA Course and Suggested Readings

NVIDIA Course

- Accelerating End-to-End Data Science Workflows

Suggested Readings

- Scaling Data Pipelines: AT&T Optimizes Speed, Cost, and Efficiency With GPUs
- Best Practices for Multi-GPU Data Analysis Using RAPIDS With Dask
- RAPIDS: Learn More
- NVIDIA AI Workbench Data Science Quickstart

Descriptive Analysis and Visualization – 13%

- Exploratory data analysis (EDA) and descriptive statistics
- Visualization
- Selecting appropriate plots for different analysis goals
- Hypothesis testing and statistical significance evaluation
- Interpreting patterns, trends, and relationships in data

Recommended NVIDIA Course and Suggested Readings

RNVIDIA Course

- Accelerating End-to-End Data Science Workflows

Suggested Readings

- Guide to Data Visualization With RAPIDS
- Making a Plotly Dash Census Viz Powered by RAPIDS
- pandas Documentation
- Top Tools for Data Exploration and Visualization With Their Pros and Cons
- Accelerated Data Analytics: Speed Up Data Exploration With RAPIDS cuDF

Foundations of Accelerated Data Science – 12%

- Python fundamentals for data analysis (NumPy, pandas, Jupyter)
- Core GPU acceleration concepts and advantages for data science
- CPU versus GPU workloads and memory transfer optimization
- End-to-end data science workflow (ingest, ETL, clean, transform)
- Distributed versus GPU-accelerated computing frameworks
- Model parameters, tuning, and overfitting versus underfitting concepts

Recommended NVIDIA Course and Suggested Readings

RNVIDIA Course

- Accelerating End-to-End Data Science Workflows

Suggested Readings

- Data Science With Python Tutorial—GeeksforGeeks
- RAPIDS cuDF: GPU-Accelerated DataFrames
- RAPIDS cuDF Versus pandas Comparison
- What is Accelerated Computing?
- Supervised Versus Unsupervised Learning

Introductory MLOps Practices – 10%

- Monitoring and optimizing ML pipelines for performance and reliability
- Managing and tracking experiments with MLflow, Weights & Biases, or custom tools
- Model saving, loading, and prediction generation
- Monitoring production models for drift and performance degradation
- Managing model artifacts and configurations for reproducibility
- Benchmarking workflows and selecting optimal hardware

Recommended NVIDIA Course and Suggested Readings

RNVIDIA Course

- Accelerating End-to-End Data Science Workflows

Suggested Readings

- A Guide to Monitoring Machine Learning Models in Production
- Prometheus Histograms Best Practices
- ML Model Monitoring in Production: Best Practices (Datadog)
- MLflow: Hyperparameter Tuning and Experiment Tracking
- Mastering Version Control for ML Models (DagsHub)

Advanced Data Structures – 7%

- Time series data handling, splitting, and forecasting evaluation
- Managing missing or irregular timestamps with cuDF interpolation
- Graph-based data representation and analysis
- Node importance evaluation and network relationship visualization

Recommended NVIDIA Course and Suggested Readings

RNVIDIA Course

- Accelerating End-to-End Data Science Workflows
- Accelerated Machine Learning for Time Series Forecasting

Suggested Readings

- Accelerating Time Series Forecasting With RAPIDS cuML
- Difference Between Graph and Relational Databases (AWS)
- Beginner's Guide to GPU Accelerated Graph Analytics in Python | NVIDIA Technical Blog

Software and Environment Management – 6%

- Enable reproducibility in data science projects by maintaining environment files
- Configure reproducible Python environments using Conda, pip, or Docker
- Efficiently manage software dependencies and collaborate in multi-user data science environments
- Perform GPU environment check (driver/CUDA®/RAPIDS compatibility, nvidia-smi, device visibility) and resolve a dependency conflict
- Understand the basics of version control using Git

Recommended NVIDIA Course and Suggested Readings

RNVIDIA Course

- Accelerating End-to-End Data Science Workflows

Suggested Readings

- Get Started With GPU Acceleration for Data Science
- NVIDIA SMI User Guide
- Conda Environments Guide
- NVIDIA Container Toolkit—Docker Integration
- Basic Git Commands (Bitbucket Server Docs)

Questions?

Contact us here.

NVIDIA