What's New in Neo4j Graph Data Science

October 2022

What Is Neo4j Graph Data Science?

Neo4j Graph Data Science is an analytics and modeling engine that uses the relationships in your data to improve predictions. It plugs into enterprise data ecosystems so you can get more data science projects into production quickly. Using pretuned graph algorithms, data scientists explore billions of data points in seconds to identify hidden connections and generate compelling visualizations that lead to better stakeholder decision making.

Areas of Investment

Neo4j offers the only graph data science engine built for data scientists to improve their predictions and ML models, at scale, with seamless integration across the data stack. We continue to build on the momentum of our 2.0 and 2.1 releases and focus on building the most comprehensive graph data science solution on the market.

We're investing in four key areas:

- **Built for data scientists:** Work in a familiar environment and quickly demonstrate practical business value.
- **Make better predictions:** Build a proof of concept and go to production using any data source to discover what's important, what's unusual, and what's next.
- **Integrate with your data ecosystem:** Integrate Graph Data Science with the existing tools across your technology stack and data pipeline using native connectors.
- **Production ready - Trusted, scalable, and robust:** Deployment flexibility and options for moving models into production get more data science projects adopted.

What's New?

Highlights from this release include:

Built for Data Scientists

Work in a familiar environment and quickly demonstrate practical business value with:

- **Python client improvements:**
  - A new manual and database argument handling input make it easy to find the information needed and reduce the amount of Python code required to make new database connections.

Improve Predictions

Features that enable you to build a proof of concept and go to production using any data source to discover what's important, what's unusual, and what's next include:

- **Machine learning (ML) pipeline improvements:**
  - **Expand features in ML Pipelines:** Configure *contextNodeLabels* and *contextRelationshipTypes* to generate more features from additional node labels and relationship types for a more robust and diverse set of graph features in your ML Pipelines.

  - **Configure source and target node labels in Link Prediction Pipelines:** Predict more specific kinds of relationships and minimize the number of node pair comparisons for more efficient training and prediction.

- **Graph native deep learning with multilayer perceptrons (MLPs):** In addition to linear regression and random forest, you can now choose and configure MLPs, a type of feed forward neural network, as a model candidate in Link Prediction and Node Classification pipelines.

- **Graph sampling:** Train models faster and reduce compute intensity by extracting representative subgraphs to train inductive models (GraphSAGE, link prediction, and node classification) on small, relevant datasets that can be applied to larger datasets.

- **K-Means Clustering algorithm improvements:**
  - **Improve explainability and result analysis.** Evaluate the quality and accuracy, and compare performance across different values for K:
    - Centroid and node-distance-to-centroid metrics are included in the results.

    - Compute silhouette related metrics to help validate the quality and consistency of clusters by enabling *computeSilhouette*.

  - **Tune K-Means and maintain consistency across runs:**
    - Manually choose centroids or maintain consistency across runs using the *seedCentroids* parameter.

    - Select different sampling strategies for picking the first centroids with the *initialSampler* parameter.

      - The *K-means++* initialization option can be enabled by setting *initialSampler='kmeans++'*.

    - Use the *numberOfRestarts* parameter to run K-Means multiple times with different initialization points so you can then choose the best performing one.

Integrate with Your Data Ecosystem

New and improved connectors, extensions, and integrations across the data pipeline ecosystem include:

- **Graph topology export:** Use Neo4j Graph Data Science projections in other graph ML libraries like Deep Graph Library (DGL), PyG, and Tensorflow GNN. Combine graph topology export with graph sampling to scale graph neural networks (GNNs) to larger graphs and deploy models trained using open source ML libraries in your Neo4j Graph Data Science workflows.

Production Ready: Trusted, Scalable, and Robust

Features that improve deployment flexibility and move models in production faster:

- **Enhanced embedding performance:** Embed graphs up to 200% faster with logistic Link Prediction algorithms and 62% faster with Node2Vec algorithms with Neo4j Graph Data Science 2.2 compared to Graph Data Science 2.1.

- **Global admin configuration:** Admins can now provision storage and compute requirements for each user so that teams can run several projections concurrently, no matter their location or working time zone.

Previous Release Announcements

Neo4j Graph Data Science 2.1 - June 2022

Neo4j Graph Data Science 2.0 and AuraDS - April 2022