What's New in Neo4j Graph Data Science

June 2022

What Is Neo4j Graph Data Science?

Neo4j Graph Data Science enables fast, actionable insights about what's important, what's unusual, and what's next through an easy to use engine that works with the data you already have, the data stack you already have, and the data pipeline you already have to quickly move more data science projects from proof of concept to production.

Areas of Investment

We continue to build on the momentum of our 2.0 release highlights. Our investments focus on building the most comprehensive graph data science solution on the market.

We're investing in four key areas:

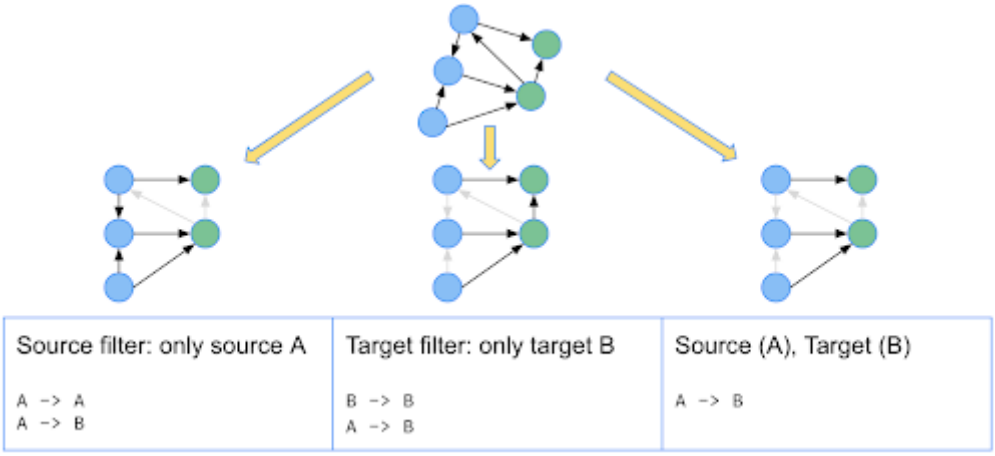
- **Easy to use:** Take the fastest path to production.
- **Graph built for data scientists:** Analyze graphs through a library of graph algorithms, ML pipelines, and data science methods.
- **Enterprise ready: trusted, scalable, and robust:** Securely handle hundreds of billions of nodes and relationships.
- **Data pipeline ecosystem:** Integrate Graph Data Science with the existing tools across your technology stack and data pipeline.

What's new?

Easy to use

Improvement of experiences to make them simple, fast, and intuitive include:

1. **Autotuning:** ML pipelines (nodeClassification, nodeRegression, linkPrediction) now support automated tuning for hyperparameters; users configure the system, Neo4j Graph Data Science finds the best parameter combinations to . provide the best performing models possible.
2. **Source & Target filtering for KNN and Node Similarity:** Similarity algorithms are some of the most popular, but often users do not need to compare every possible pair of nodes in their graph. Source and target filtering lets users limit the scope of similarity calculations to just the relevant nodes for each use case.

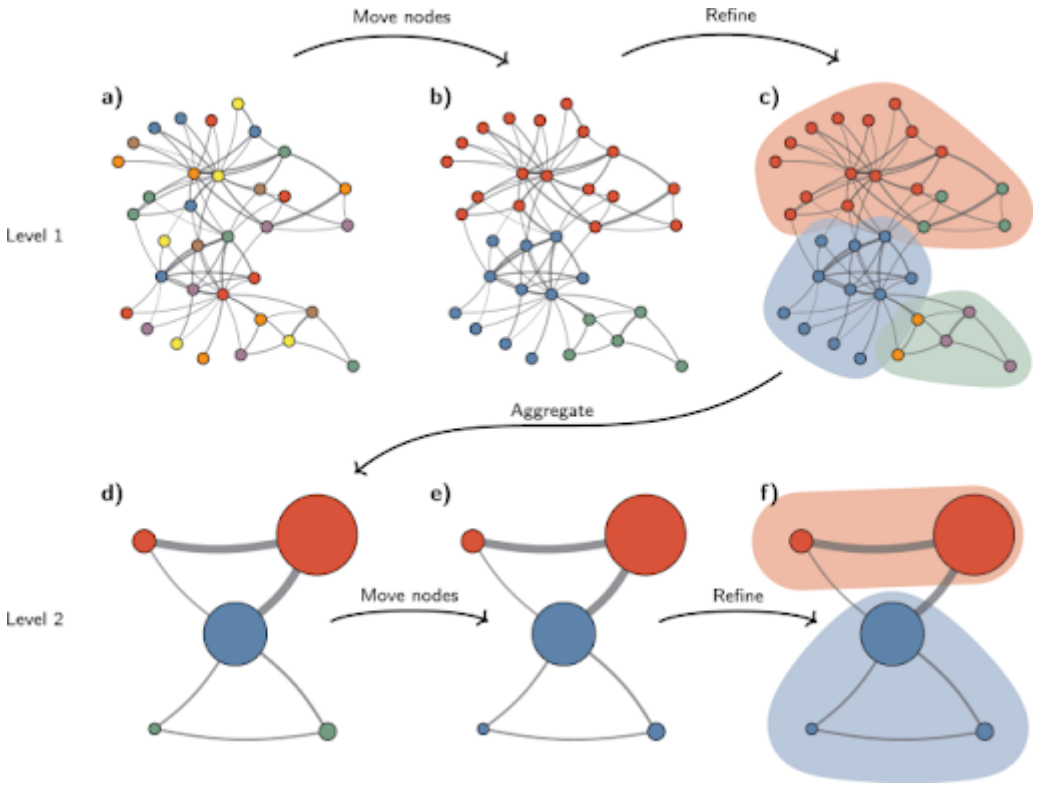


3. **Visual Progress Logging in the Graph Data Science Python Client:** now, when users run algorithms or project graphs, a progress bar is displayed that shows the status of tasks.

Graph built for data scientists

Features that empower data scientists to analyze graph are:

1. **New alpha tier algorithm** - Leiden - new community detection algorithm, a hierarchical clustering algorithm that guarantees well connected communities. Similar to Louvain, users have requested this methodology to create more cohesive communities.



From Traag, V.A., Waltman, L. & van Eck, N.J. "From Louvain to Leiden: guaranteeing well-connected communities." Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

2. **New alpha tier algorithm** - K-means clustering: community detection algorithm intended to cluster nodes based on properties (like embeddings). Users can specify the numbers of clusters desired and Graph Data Science finds the optimal groupings.
3. **New alpha tier ML pipeline** - Node Regression: users can predict numerical property values for nodes using node regression pipelines. Node regression lets users fill in missing property values based on other node properties and graph topology.

Enterprise ready: trusted, scalable, and robust

Features that support enterprise volumes of data, use cases, and complexity include:

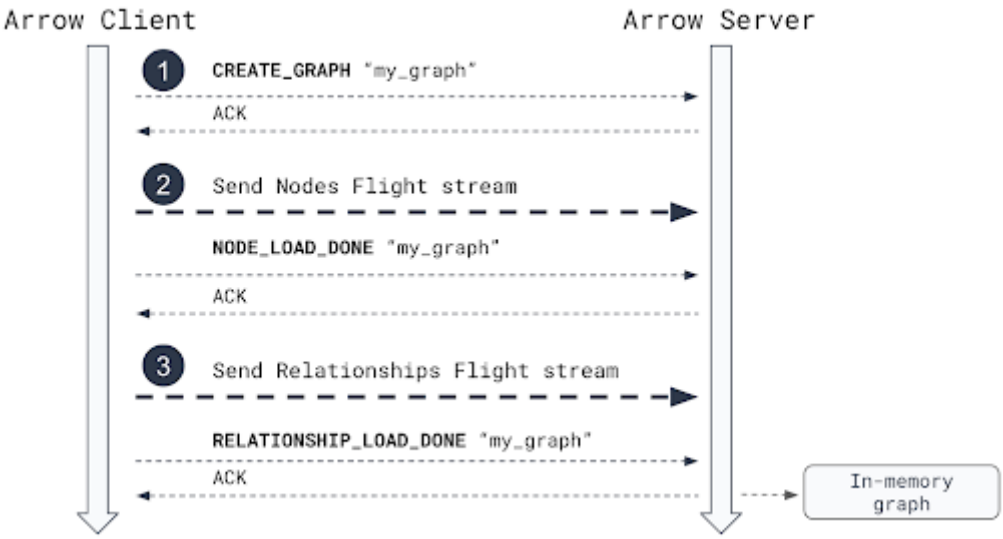
1. **Apache Arrow Integration for Graph Projections**: import and export massive graphs directly into Graph Data Science- at speeds of up to **8 million objects/second**.

   Leveraging Arrow to directly build graph connections makes it simple to Insert Graph Data Science seamlessly into your existing ML pipelines and run analytics that need to be exported to a downstream system.

   The Neo4j Graph Data Science Arrow integration provides: a built-in Arrow flight server, bundled with Graph Data Science, Arrow convenience functions in the Graph Data Science Python Client to load from and export to data frames, and access to a low level Arrow API to integrate with any Apache Arrow supported product like Google BigQuery, Beam, Parquet files, etc.

   Apache Arrow integration for graph projections is available to Graph Data Science Enterprise Edition customers only.



2. **Performance Improvements for Machine Learning**: Through optimization of internal machine learning code, the training time for GraphSAGE embeddings is up to 90% faster, Random Forest model training is up to 80% faster, and Logistic Regression is up to 40% faster.

Data ecosystem

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

1. Graph Data Science Python Client Improvements: The Graph Data Science Python Client can automatically use Apache Arrow for data movement on Enterprise licensed instances. Users can now specify the return format of data frames when streaming node properties or relationship results. The Graph Data Science Python Client supports all Graph Data Science 2.1 features.
2. Neo4j Data Warehouse Connector offers a simple way to move data between the Neo4j database and data warehouses like Snowflake, Google BigQuery, Amazon Redshift, or Microsoft Azure Synapse Analytics. It can be used as a Spark Submit Job (by providing a JSON configuration), or with a Scala/Python API that simplifies writing the Spark job to move data between the Neo4j database and the data warehouse.

Previous Release Announcements

1. Neo4j Graph Data Science 2.0 and AuraDS - April 2022