4 WAYS KNOWLEDGE GRAPHS ACCELERATE DATA SCIENCE FOR THE ENTERPRISE

Introduction

In today's business climate, organizations must readily adapt to change and uncertainty. Data-driven automation, in the form of machine learning and artificial intelligence (AI), is necessary for today's enterprises to both survive and thrive.

Data landscapes are shifting towards AI-infused automation – and organizations can't afford to fall behind fast-moving competitors. Today, data science has two main challenges:

- Data sources take too long to access, extract, and clean.
- DevOps teams struggle in deployment, meaning teams are grappling with operationalizing individual data scientists' machine learning models.

Semantic graph platforms are uniquely well-suited for data science projects. They make it extremely easy for data analysts to blend additional data sets into harmonized data collections. Graph platforms expose and express relationships between data points. In many cases, they enable support for generating supplemental connections through inferences and algorithmically produced data linkages. Many functions help expose and clarify signals needed for machine learning.

A scalable knowledge graph platform with these capabilities makes data science initiatives faster, simpler to operationalize, and more effective. Knowledge graphs support fundamental data science processes such as data onboarding, integrating, blending, and engineering for machine learning features. Additionally, knowledge graph platforms integrate with the most popular data science tools available, producing two crucial results:

- Data wrangling and preparation burdens are minimized when provisioning data for data science projects. This reduces the time needed to build new analytic-ready datasets.
- Tightly integrated, connected, and multidimensional data is delivered. This can better reveal insights through the application of machine learning techniques.

1. Data Preparation

Data preparation steps associated with data science processes demonstrate how knowledge graph solutions eliminate bottlenecks and speed data delivery. Initial processes often require professionals to prepare data for feature engineering or identify data characteristics that increase models' predictive accuracy. For example, a model that detects similarities across products must be trained on many attributes. This includes their intended function, dimensions, construction materials, operational features, and more.

Knowledge graph technology offers an unmatched ease of integrating multiple data sources for feature engineering, including data sources with differing attributes. Data is conformed to a standardized semantic conceptual model that tames complex data representations, regardless of original differences in schema, structural variation, units of measure, or other discrepancies between different data. Knowledge graph technology, such as Altair® Graph Studio™, accomplishes this task with automated transformation pipelines that reduce the time spent on integrating data among sources. This function maximizes the value of data preparation by quickly assembling greater numbers of data sources. It tightly connects and integrates data sources to expose additional signals that could not have been gleaned from any sole source. Blended datasets are then published to any desired analytics or machine learning tool.

2. Feature Generation

Algorithms

The formal feature generation process begins once data is accessed, integrated, and blended during data preparation. Compelling solutions in the semantic graph realm contain several mechanisms for optimizing feature identification by pinpointing signals in data. One useful solution is a series of algorithms that generate data about the graph for analytics and denoting features. These algorithms are specific to graphs and data science, respectively. Graph algorithms include techniques such clustering, nearest neighbors, and PageRank. For example, clustering might illustrate which customer characteristics – such as customer service interactions or changes in income, address, or job – have the greatest impact on churn.

More common data science algorithms include principal component analysis (PCA), singular value decomposition (SVD), and several other methods that support exploring features to better understand datasets. PCA, for example, can show a data scientist how properties of interest (such as the results of clinical trials) are or are not correlated. This provides insights on suitability for selection as an individual feature. Other algorithms for dimensionality reduction and classifiers also refine and prepare data for analyzing features.

Automation

Data exploration has two types of automation: automatic query generation and automatic data profiling. The former enables rapid feature engineering on data combinations. Scientists build tabular extractions from the knowledge graph without time-consuming coding. They are modeling part of a feature with each column output. An output table is built from the graph and fed downstream to machine learning tools. The machine learning tool will be trained to create predictions for one of its columns. Each column includes data from various parts of the graph, which may have been transformed (action, etc.) by the extraction query. This task is expedited with automated graph query generation. Data scientists create wide tables with column transformations applied in the feature set. This contributes to the overall extraction graph query. Without it, data scientists must write code to isolate each feature. The amount of feature experimentation would be either minimized or prolonged.

The ability to automatically profile data is pivotal for quickly revealing types of data populating graphs. When many sources of data are integrated, the resulting graphs can become overwhelming. Tools are needed to understand what data is available and how it's linked. Auto-profiling delivers detailed statistics of every dimension of the data that can be visualized. This reveals its potential use for feature identification. This task is quicker with automation, making it easier to build the feature tables. Users can extend the automated results with their own profiling queries and add into the system any profile type of interest for future efforts.

3. Building Models

Data science processes fluctuate during stages of refining models in production. After data prep and feature engineering, the process involves training predictive models. Knowledge graph technology enhances this phase in two ways, by:

- **Seamlessly integrating with premier big data and advanced analytics machine learning ecosystems**: There's no shortage of data science and machine learning tools for building predictive models, such as Altair® RapidMiner®. Altair's machine learning solutions allow users to easily train, evaluate, explain, and deploy predictive and prescriptive models within a low- or no-code environment.
- **Supporting geospatial functions**: A complete knowledge graph toolset should provide a suite of geospatial functions. These functions can be useful depending on an organization's domain or advanced analytics use case. For instance, machine learning systems used to improve shipping logistics or predict the effects of delays might benefit from geospatial capabilities.

These capabilities provide foundational support for constructing and deploying machine learning models with options supporting an array of use cases and data science needs.

Operationalizing Models

A frequent data science challenge is successfully deploying models into production and operating workflows. Knowledge graph technology addresses this need by making APIs available for user-defined functions (UDF), user-defined services (UDS), and the graph-data interface (GDI). For example, in Graph Studio, these gateways enable capable semantic graph solutions to push data to third-party tools and apply those models to data integrated in the graph. This includes inferring new graph data, countering model drift, refreshing models, and refining models with new features and data based on production results. These techniques are crucial for ensuring models scale appropriately and involve:

- **User-Defined Function (UDF)**: This API enables organizations to use external models available through REST services to augment graph data. Data scientists may want to leverage models in Altair RapidMiner's scoring agent or in Amazon SageMaker. For example, graph data about products can be analyzed when developing new features. UDF calls to external models, passing them through part of the graph's data (the equivalent of a single record); in return, UDF receives a prediction to add to the graph. This shows which feature would perform best for customers. It's a method for applying new or existing graph data to machine learning models.
- **User Defined Services (UDS)**: This API mechanism imports and exports data into external tools or graphs. UDS enables data scientists to use any code (such as C++ or Java-based languages) for integrating with data science tools. The close integration with the underlying semantic graph solution and UDS leverages the latter's parallel processing capabilities for quick data imports and exports. Parallel data processing speeds up data science tasks across all CPUs in each server node. The platform's cluster executes these queries.
- **Graph Data Interface (GDI)**: The GDI is a subsystem used for managing integration with and connectivity to myriad upstream and downstream data systems for moving data in and out of the graph using queries. The GDI supports pushing and pulling data from HTTP REST services for analytics tools without programming. It offers virtualization capabilities as an alternative to direct data importation. Users can leave data where it is and still access it as though they're in the graph.
- **Data Layers**: Data layers are metadata-defined pipelines that can be used for creating, testing, and operationalizing machine learning models. This framework can create and orchestrate sophisticated multi-query pipelines with third-party systems. This includes production systems for updating models to ensure they're performing as they do in training environments. Data layers are used for orchestrating the integration of data for analytics.

Each of these mechanisms enables organizations to go beyond simply preparing data for creating and training machine learning models by empowering them to operationalize them. This enables enterprises to cover the entire process of applying machine learning and broaden its business value.

4. Graph as a Better Data Representation for Machine Learning

Graphs accelerate and improve data science in production through their flexible, uniform data representation. Machine learning models augment data in the graph to create inferred predictions. Users can add metadata information, including which metadata was used for each prediction and when. Metadata also includes contextual information, such as prediction confidence scores or records of actual model performance.

Knowledge graph technology accelerates and improves data science. It speeds up time-consuming data preparation and feature engineering while solidifying and scaling efforts in production. These advantages combine with more features: incorporating labeled properties, embedding, and efficiently using a data science tool to provide a solid foundation for enterprise automation when it is most crucial.

About RapidMiner

Altair RapidMiner, our data analytics and AI platform, helps you overcome the most challenging obstacles in your data journey. We offer a path to modernization for established data analytics teams, as well as a path to automation for teams just starting their data journey.

Altair's team of technical and commercial experts assist organizations with their unique journey to becoming a truly AI-driven enterprise. To learn more about how knowledge graphs can help make your data easier to use, scalable, and advance your analytics insights, please reach out for a consultation at https://altair.com/contact-us.