SOLUTIONS FLYER

USE CASE: INTEGRATING UNSTRUCTURED DATA SOURCES INTO AN ENTERPRISE KNOWLEDGE GRAPH

Modern enterprise data ecosystems store information as textual data, so an enterprise-scale knowledge graph platform must be capable of integrating large collections of unstructured data. Knowledge graphs created by Altair® Graph Studio™ connect to the source's facts to analyze them alongside more structured data, like relational databases, data lakes, or application programming interfaces (APIs). Graph Studio onboards unstructured data directly into knowledge graphs using a configurable, scalable pipeline that requires no customized coding. These pipelines generate a graph model for unstructured text and extracted metadata. This creates connections between elements and related entities for fully integrated search queries and semantic relations within the graph.

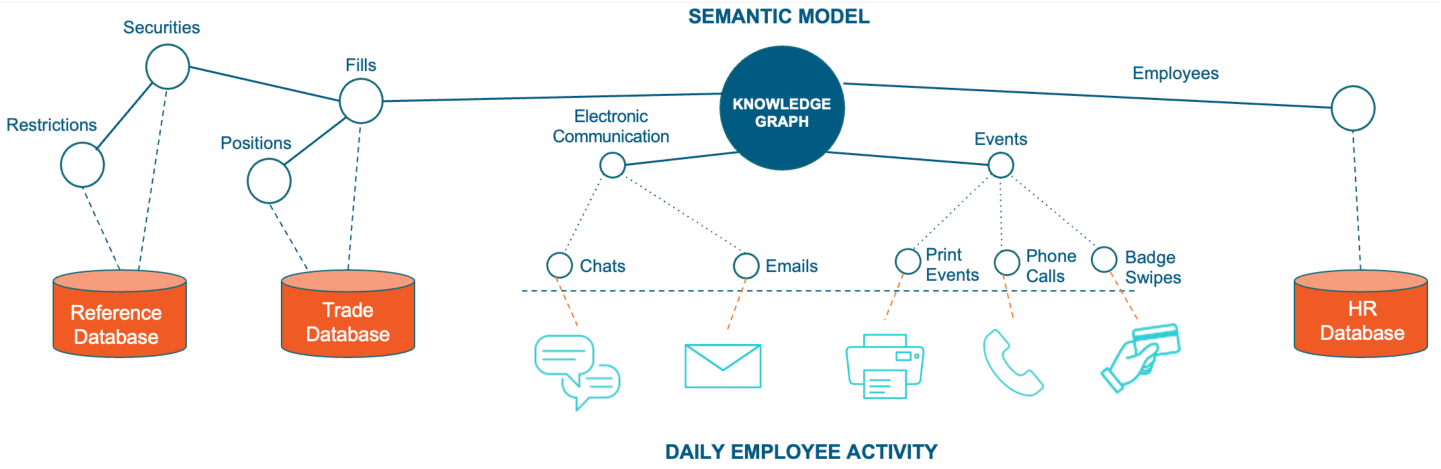
Through the following example, we'll demonstrate how Graph Studio inputs unstructured data — sources that contain text, like PDFs, text messages, or text snippets embedded in structured data — directly into knowledge graphs.

Context and Requirements

A large company needs to build an integrated, large-scale search application that leverages data from multiple applications across the enterprise. Unstructured and structured data from previously siloed applications must be harmonized into a single, flexible model. Graph Studio provides a comprehensive, scalable knowledge graph that brings these data sources together.

There are complex, technical requirements the knowledge graph must consider. It must be able to:

- Allow analysts to execute text-based searches across millions of documents, going back years in the past, with interactive (1-3 second) response times.
- Accommodate hundreds of thousands of new documents added to the knowledge base daily, searchable within hours of the documents becoming available.
- Enable complex filtering and sorting of text-based searches with additional criteria of related but distinct metadata attributes, e.g. people, dates, document classification categories.
- Support the development and use of a flexible data model (ontology) that harmonizes structured and unstructured data and can be updated over time.
- Incorporate and surface results from a cloud-based, machine learning-driven text analytics engine used for document classification and facilitate the development, testing, and validation of this analytics engine to improve its performance over time.



Implementation

Onboarding

- Every day, Graph Studio's pipelines crawl hundreds of thousands of documents into a production knowledge graph.
- Graph Studio's pipelines send text from each document to an endpoint that hosts the text analytics engine. The endpoint provides structured information about the documents (output). From there, the output is incorporated into a knowledge graph, connecting related entities.
- Graph Studio's pipeline builds an ElasticSearch text index that indexes the text of communication records for searches through knowledge graph queries.
- A real-time pipeline is scaled to process thousands of documents per minute, every day, with no human intervention required.
- The user's knowledge graph is populated with document references, text contents, and structured data, directly linking related entities in the knowledge graph.

Querying and Analysis

- As a result of the onboarding process, an enterprise knowledge graph is enriched with output from three years of unstructured text content, enabling ElasticSearch to conduct a querying and analysis experience.
- Customer analysts leverage the platform to quickly and easily answer complex questions.
- Using a tailored front end that sits on top of the text-enriched knowledge graph, analysts conduct free text searches across millions of documents stored within the knowledge graph. Searches are filtered based on connections with the harmonized structured data.
- The searches and other UI controls are translated to queries and executed against Graph Studio, which combines ElasticSearch text querying alongside SPARQL queries that traverse multiple hops in the knowledge graph.
- Leveraging the processing power of hundreds of CPU cores operating in parallel in the knowledge graph MPP query engine cluster and ElasticSearch cluster, most queries return results in under two seconds.
- The end user experiences a dynamic, interactive search platform powered by free text searches coupled with complex queries across different data dimensions within the knowledge graph.

Text Analytics Engine Development

- Graph Studio's knowledge graph platform and unstructured data integration are used as a flexible validation framework for the development of text analytics engines.
- In a pre-production environment, the data science team uses Graph Studio's unstructured data pipelines to rapidly feed large volumes of historical unstructured data as input into their text analytics engine.
- Graph Studio uses the engine's outputs in the knowledge graph for easy analysis with the related structured data.
- The data science team leverages the non-production pipeline as the basis of a powerful back-testing and validation framework.
- The team iteratively runs historical documents through the pipeline, adjusting the classification model between runs.
- They analyze the outputs alongside each other in the graph and compare precision-recall thresholds across runs.
- Graph Studio's scalable onboarding process and in-memory query engine reduce operational and financial costs.
- The **result** is a comprehensive, agile workflow that enables iterative development of the text classification model, driven by and fully integrated with the customer knowledge graph.