ALTAIR KNOWLEDGE GRAPHS: ARCHITECTED FOR THE ENTERPRISE

Large enterprises struggle integrating siloed data across thousands of sources, often requiring months of custom extract-transform-load (ETL) development for each new business question. Traditional data approaches create brittle architectures that resist change, making it difficult to derive value from organizational data assets.

Altair® Graph Studio™, part of the Altair® RapidMiner® data analytics and artificial intelligence (AI) platform, addresses these challenges by helping organizations create, manage, and leverage knowledge graphs at scale. The software connects to disparate sources of structured, semi-structured, and unstructured data and uses a semantic layer approach to data integration. We developed Graph Studio for large global enterprises who need effective ways to integrate and query their data and receive fast, accurate, and complete responses.

THE KNOWLEDGE GRAPH ADVANTAGE: INVERT THE DATA INTEGRATION PARADIGM

Traditional data warehousing and data lake approaches require extensive upfront definition of business questions, precise schema design, and complex ETL pipelines. This results in brittle architectures optimized only for predefined queries; adding new data sources or answering new questions is usually expensive and time-consuming.

Altair's approach to knowledge graphs inverts this paradigm to improve the scalability of integration and enhances the flexibility of knowledge graph-enabled data analytics systems.

The traditional data warehouse approaches:

- Require extensive upfront definition of business questions and reports.
- Demand precise schema design based on anticipated queries.
- Involve creating complex ETL pipelines targeted at specific outcomes.
- Result in brittle architectures optimized only for predefined questions.
- Make adding new data sources or answering new questions expensive and time-consuming.

> By modeling the semantic relationships between data elements rather than predicting specific queries, our graph technologies enable organizations to blend diverse data assets—structured and unstructured, internal and external—into a flexible, adaptable fabric that grows with business needs.
>
> Sam Mahalingam, CTO, Altair

Altair Graph Studio is an enterprise-scale knowledge graph toolset that enables agile data integration, transformation, and analytics for organizations with diverse data sources.

In contrast, Altair's knowledge graph approach supports:

- **Domain-centric modeling:** Focus on modeling the domain, its concepts, and their relationships rather than addressing specific reporting or analytics queries.
- **Ontology-based integration:** Automatically map available data sources to a logical/conceptual ontology model, providing a semantic layer in the process.
- **SPARQL queries:** Allow users to ask any question without requiring analysts to predict likely queries in advance.
- **Seamless extension:** Add new data sources through a single graph edge connection to existing knowledge graphs without disrupting existing functionality.
- **Query preservation:** Maintain all existing queries, reports, and dashboards when adding new data sources.
- **Complex relationships and flexible shaping of data:** Natively represent hierarchies, cyclical relationships, multiple inheritances, dynamic attribute typing, and classifications.
- **Universal integration:** Integrates structured and unstructured data sources.

Graph Studio overlays a semantic, graph-based data fabric on diverse enterprise data sources with ontologies that add real-world business meaning to structured and unstructured data, accommodate new sources with minimal impact on existing implementations, and enable a single pane of glass for enterprise data.

Knowledge graphs store relationships between information points sourced from disparate repositories in virtually any data format. They facilitate efficient query responses, unify diverse data, improve data quality, and eliminate hallucinations when used with large language models (LLMs).

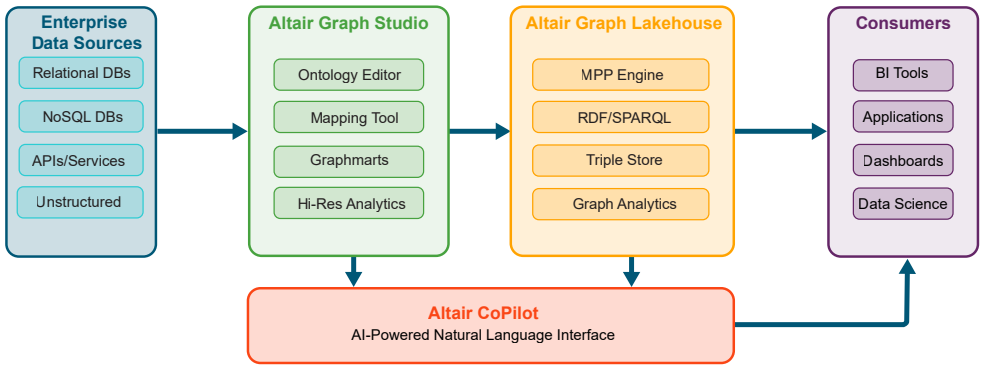
PLATFORM ARCHITECTURE OVERVIEW

Altair's knowledge graph solution consists of three integrated modules:

- **Graph Studio:** A comprehensive knowledge graph tool that integrates data from structured and unstructured sources and provides tools to create ontologies, map data, and generate dashboards.
- **Altair® Graph Lakehouse™:** A massively parallel processing (MPP) graph database engine designed for high-performance, data-intensive workloads.
- **Altair® CoPilot™:** A conversational business intelligence tool that leverages generative AI (genAI) to enable natural language interaction with graph data.

These three components work in concert: Graph Studio provides the data fabric layer and semantic modeling capabilities as instructions that drive Graph Lakehouse, which delivers the computational power for complex transformations and analytics. Altair CoPilot makes all this sophisticated analytics technology accessible through natural language.

Altair Graph Technologies Platform Architecture



Knowledge graphs built with Graph Studio provide understandable, comprehensive organizational context to enterprise analytics solutions, including those employing genAI.

> Altair's massively parallel architecture transforms what's possible with knowledge graphs. It scales to billions of data points while maintaining performance integrity, proving that enterprises no longer need to choose between comprehensive data relationships and analytical scale.
>
> Sam Mahalingam, CTO, Altair

KEY DIFFERENTIATORS

Altair's graph technologies feature several unique aspects that differentiate them from traditional databases and other knowledge graph tools:

Multi-Graph Architecture

- Maintains separate subgraph contexts while enabling cross-subgraph analytics.
- Provides an agile, iterative approach to data integration.
- Supports the creation of nonlinear ingestion and transformation workflows.
- Allows data engineers to swap variations of transformation pipeline steps instantly and iterate quickly.
- Enables fine-grained access control for specific subgraphs.

MPP Graph Lakehouse Engine Design Philosophy

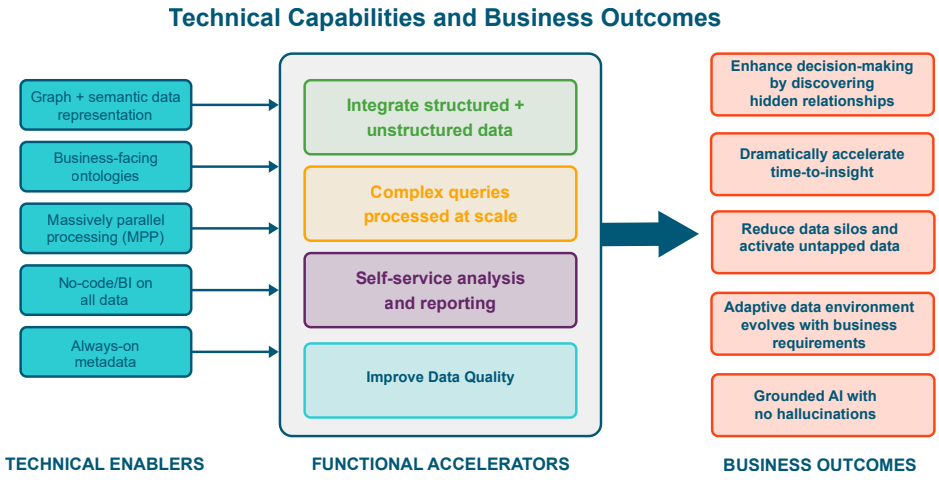
- Prioritizes throughput over latency.
- Utilizes all available resources all the time for maximum query efficiency.
- Optimized for analytical workloads processing billions of data points simultaneously.
- Shards triples across nodes and distributes queries in a fully parallel fashion.

Schema-less Design and Dynamic Modeling

- Allows ingestion of any data shape and extraction of every bit of information, including imperfectly structured or "dirty" data.
- Dynamically creates ontology models and transformation mappings for upstream data sources.
- Supports in-graph data cleaning and standardization.
- Eliminates rigid schemas that limit adaptability.

Comprehensive View Support

- Provides both dynamic and materialized views, a unique feature among graph databases.
- Enables sophisticated data presentation without modifying underlying data.
- Supports virtual views across multiple source systems.



Graph Studio identifies the location, content, and contextual business meaning for all your data. It exposes connections between datasets and enables rapid visual data exploration and discovery. Graph Studio's unique use of semantics and a graph data model in easy-to-use tools enable engineers, managers, citizen data scientists, and business analysts to explore enterprise data with ad hoc and/or predefined questions, blend data across previously siloed platforms, and build analytics-ready datasets in minutes.

> The semantic foundation provided by knowledge graphs built with Graph Studio improve data discovery and analytics and also provides critical context for genAI, eliminating hallucinations and improving response quality.
>
> Sam Mahalingam, CTO, Altair

DATA MANAGEMENT LAYER

Knowledge Graph Development Tools

Graph Studio enables data engineers and domain experts to build automated pipelines that integrate all available data into knowledge graphs, support complex ad-hoc queries, and create sophisticated AI and machine learning workflows.

It blends data from different sources into consistent graph models accurately and reliably. Because the data is structured as a knowledge graph, users can execute ad hoc business intelligence queries and utilize innovative analytical techniques like network analysis to improve insights in a self-service environment.

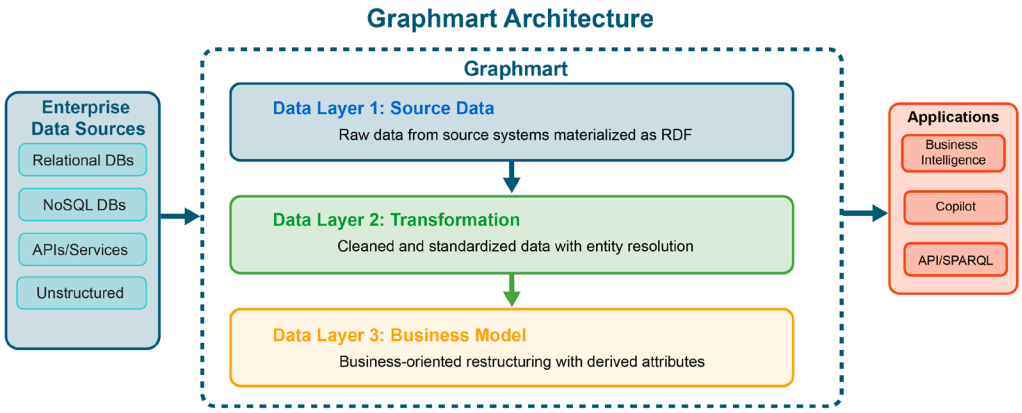
Graph Studio consists of:

- Graphmarts: Central organizational tooling for creating and managing data products and integration pipelines.
- Ontology editor: Tools for managing semantic models and relationships.
- Mapping tool: Table-style interface for connecting source data to ontologies and transforming data between ontologies.
- Dataset catalog: User-friendly interface for discovering available datasets.
- Network navigator: Use the software's intuitive visual interface to graphically explore the knowledge graph.

GRAPHMARTS: ENTERPRISE DATA PRODUCTS

Enterprises can define hundreds or even thousands of graphmarts, and each one can transform and shape disparate data to solve specific business problems. Each graphmart is a comprehensive metadata artifact describing everything required to deploy a knowledge graph-based data product, including:

- Virtual hardware configuration accessible with Kubernetes.
- Data load, virtualization, and transformation steps that drive Graph Lakehouse.
- Comprehensive data profiling metrics metadata.
- OData REST endpoint definitions.
- Ontologies providing the semantic layer.
- Associated dashboards and visualization configurations.



Graphmarts are the primary mechanism in Graph Studio for delivering knowledge graph-based capabilities to users.

> Integrating Graph Studio allows organizations to create a sturdy foundation for their advanced analytics ecosystems. It gives them the tools they need to create a true organizational data fabric – all within a single, unified, scalable, and easy-to-use platform that works with the investments they have already made in their data infrastructure.
>
> Sam Mahalingam, CTO, Altair

DATA INTEGRATION APPROACHES

Graph Studio brings together the best elements of a data fabric and an AI platform to create an enterprise AI fabric. It exposes the data — and the relationships between data objects and concepts — to the people who make mission-critical decisions.

Data Materialization (ETL)

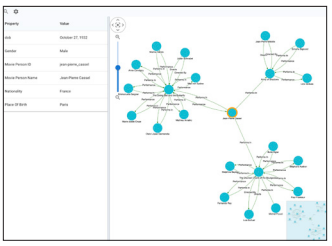
- Leverages Graph Lakehouse's MPP architecture for parallel extractions of data from source systems and loading it into the knowledge graph.
- Transforms data during loading to conform to the semantic model familiar to business users.
- Creates persistent graph representations optimizing query performance.
- Supports historical analysis with point-in-time snapshots.
- Enables analytics even when source systems are unavailable.

In-Graph Transformation (ELT)

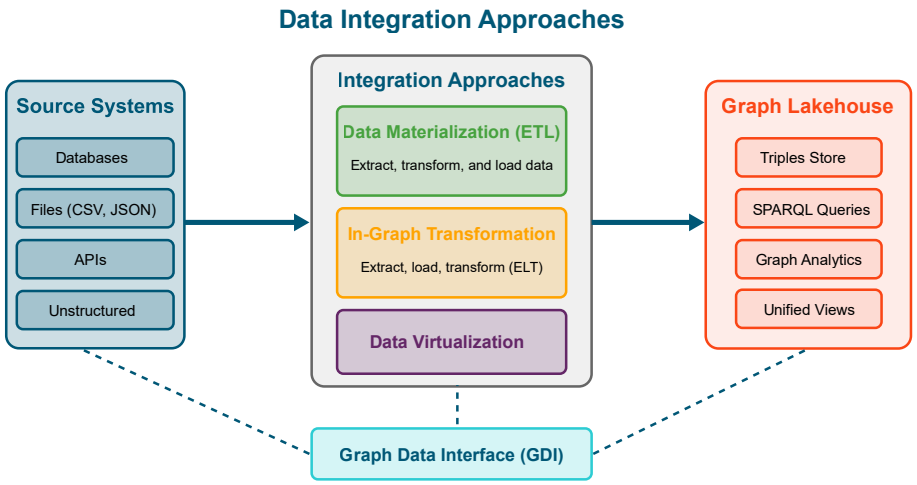
- Performs complex transformations on data already loaded into the graph.
- Leverages Graph Lakehouse's MPP architecture for transformational SPARQL queries.
- Supports ontology-to-ontology mappings with rich transformation formulas.
- Enables iterative, multistage, multidimensional transformation through data layers.
- Supports sophisticated data operations, including data cleaning and standardization, graph edge linkage between entities, deduplication and entity resolution, and business-oriented data model restructuring.

Data Virtualization

- Provides direct push-down parallel querying of source systems without data movement.
- Offers real-time access to the latest source data.
- Reduces storage requirements for large datasets.
- Preserves source system security contexts.
- Supports federated queries across multiple sources.



Graph Studio's network navigator represents each data point with a shape connected to the other data points in the graph. Users can define shapes, colors, filters, and more based conditions they select.



Graph Studio uses an in-memory MPP processing paradigm to execute queries against datasets, enabling agile data integration, transformation, and analytics at enterprise scale.

PROCESSING LAYER: MPP GRAPH DATABASE

Graph Studio includes a native MPP graph database (Graph Lakehouse) that's built to handle graph workloads for analytics and data transformation. The database can scale as needed to support any amount of data and clusters of any size.

The system utilizes:

- A "shared nothing" architecture for horizontal scalability.
- In-memory processing using compressed in-memory and on-disk storage.
- Optimized triple indexing strategy for Resource Description Framework (RDF) triple data
- Graph Online Analytical Processing (Graph OLAP) orientation.
- Real-time interactive performance for ad hoc queries with complex multi-hop joins.
- Automatic sharding of graph data across cluster nodes.
- Atomicity, Consistency, Isolation, and Durability (ACID compliance).

The graph database can operate fully in-memory for maximum performance or use a streaming disk-based mode that allows the data to remain on disk until it's needed. The system prioritizes throughput over latency and utilizes 100% of available resources for maximum efficiency on analytical workloads. For concurrent queries, resources are shared proportionally, optimizing total work-load completion rather than individual query response times.

ACCESS LAYER

Conversational Interface

Graph Studio also provides a conversational interface (Altair CoPilot) that utilizes LLMs to generate accurate graph queries and genAI for natural language interaction with enterprise data. The software:

- Uses ontological descriptions that align with textual representations to provide context.
- Processes natural language inputs to identify query intent and parameters.
- Generates appropriate SPARQL queries to retrieve relevant information.
- Presents query results in ordinary language along with data visualizations.
- Maintains context through conversational sessions.

The conversational interface uses the knowledge graph to provide ontological context to LLM models and ground genAI models. When properly implemented, this eliminates hallucinations and provides high-quality responses. Since Graph Studio uses the LLM only for the natural language interaction, responses are always fully grounded in the source data; the system simply can't hallucinate.

High Resolution Analytics and Integration

- High resolution analytics: Codeless business intelligence dashboard builder that automates the generation of complex graph queries for data exploration.
- OData endpoint builder: Seamless integration with popular data analysis and visualization tools including Altair® Panopticon™, Tableau, Microsoft® Power BI®, and Jupyter Notebooks).
- APIs and integration: Programmatic access and automation capabilities.

IMPLEMENTATION AND DEPLOYMENT

Cloud-Native Deployment

Graph Studio is ready for deployment on AWS, Azure, and GCP. It uses containerization and Kubernetes operators that may also be used on premises with Kubernetes stacks like OpenShift. Graph Studio interfaces directly to the Kubernetes API and operators that manage infrastructure life cycle. It deploys and manages instances of unstructured data pipeline clusters or Graph Lakehouse clusters dynamically to ensure scalable deployment.

Deployment Process and Implementation Timeline

With careful planning, rigorous testing, and automated execution, most clients complete their deployment in well under a year:

- Proof of concept: 4-6 weeks.
- Initial production deployment: 8-12 weeks.
- Enterprise-scale implementation: 3-6 months.

It's crucial that the process ensures a smooth transition to production and ongoing monitoring for optimal performance and stability. The major steps are:

1. Planning and requirements

- Define use cases and business objectives.
- Identify data sources and integration requirements.
- Determine hardware/cloud requirements.

2. Initial setup and configuration

- Deploy core platform components.
- Configure security and access controls.
- Establish connection to data sources.

3. Graphmart creation to support knowledge graph development

- Define graphmart data layers and integration steps.
- Automatically map, load, and transform data sources with generated ontologies into a knowledge graph.
- Automatically profile the data in the graph to understand it and expose quality issues.
- Define data quality and integration transformation workflows.
- Transform data to ontology models representing how business users prefer to see the data.
- Configure graphmart deployment parameters.
- Establish access controls and sharing settings.

4. Dashboard and analytics configuration

- Create high-resolution analytics dashboards.
- Configure OData endpoints for business intelligence tool access.
- Set up Altair CoPilot for conversational access.

5. Production deployment and management

- Implement life cycle management policies.
- Configure monitoring and alerting.
- Establish backup and recovery procedures.

Deployment Options

- Cloud deployment (AWS, Azure, GCP).
- On-premises deployment (OpenShift and other Kubernetes stacks).
- Hybrid approaches combining cloud and on prem.

Security and Access Control

Graph Studio can implement a sophisticated attribute-based access control system:

- Security policies defined and managed within Graph Studio or imported from Immuta.
- Dynamic augmentation of queries with "data firewall" inline filtering views based on policies.
- Context-aware user access control adapting to the content of the data.
- Consistent enforcement across all access methods.
- Complete audit trail of security decisions and data access.

Additional security features include:

- User identity authentication through Keycloak.
- User and group role-based access control (RBAC) to data layer subgraphs.
- Integration with enterprise directory services.
- Secret management through Hashicorp Vault.
- Audit logging and compliance reporting.

Integration Capabilities

1. External tool integration

- SPARQL endpoint exposed for programmatic access.
- User-defined OData REST interfaces for BI tools, Jupyter notebooks, and programmatic access.
- Support for standard RDF and OWL ontology import/export formats.
- Export capabilities for various file formats.
- API access for custom applications and automation.
- Web-based interfaces for embedded analytics.
- Direct to Tableau Server Hyper Export.

2. Unstructured data processing

- Natural language processing for text analysis.
- Named entity recognition and extraction.
- Relationship identification in text.
- Document classification and categorization.
- Integration of extracted entities into the knowledge graph.
- Linking between structured and unstructured data.

3. AI integration

- Integration with machine learning frameworks like Spark MLLIB, TensorFlow, and Python Pandas via Apache Arrow export/import.
- Ground generative AI models in factual organizational knowledge graph data.
- Eliminate hallucinations in AI-generated responses.
- Support for vector embeddings in RAG methods.
- Operators for direct connections from Altair® AI Studio and Altair® AI Cloud.

MAJOR BUSINESS IMPACTS

The Altair knowledge graph platform delivers significant business value through:

- **Reduced time to insight:** Slashes the time needed to prepare data for analysis, allowing business users to get answers in minutes rather than weeks.
- **Lower maintenance costs:** Eliminates the expensive cycle of redesign and rework for each new business requirement.
- **Enhanced adaptability:** Creates flexible data environments that evolve with business needs rather than constraining them.
- **Democratized data access:** Makes complex data insights accessible to business users without technical assistance.
- **Improved decision-making:** Enables users to ask unanticipated questions and receive accurate answers based on all available data.
- **Enhanced AI:** Provides critical context for genAI, eliminating hallucinations and improving response quality.
- **Open data standards:** Makes reusable datasets a reality, saving significant time and cost on future integrations.

Organizations implementing Altair's graph technologies report:

- Reports and dashboards that previously took weeks to produce are now completed in hours or minutes.
- Queries that previously took hours or were impractical now execute in seconds.
- New data sources integrate in hours and days rather than months.
- Dramatic reduction in development backlog for new business data product requirements.

TECHNICAL REQUIREMENTS

Hardware Recommendations: Development Environment

- 16+ CPU cores
- 64GB+ RAM
- SSD storage

Hardware Recommendations: Production Environment

- Scale based on data volume and query complexity
- Cloud deployment recommended for scalability

Software Prerequisites

- Kubernetes (for dynamic deployments)
- NFS or other POSIX filesystems storage for graph data

Supported Data Sources

- Relational and NoSQL databases (via JDBC)
- Elasticsearch
- REST APIs
- Structured files (CSV, XML, JSON)
- SAS Transport formats
- Unstructured text documents

PROFESSIONAL SERVICES AND NEXT STEPS

Altair offers professional services to help organizations develop and implement knowledge graphs:

- Architecture and design consulting
- Implementation services
- Data modeling and ontology development
- Knowledge graph best practices
- Training and enablement

Altair can also partner with major systems integrators on large-scale projects.

NEXT STEPS

- Schedule a technical deep dive: Connect with our solutions architects to explore how Graph Studio can address your specific data challenges. Schedule now at altair.com/contact-us.
- Define a proof of concept: Our team will help you identify a high-value use case to demonstrate the capabilities of the platform in your environment. Contact us at altair.com/contact-us.
- Access documentation and resources: Explore our comprehensive technical documentation and learning resources to prepare for implementation. Learn more at docs.cambridgesemantics.com.