DataWalk

DataWalk: Introduction for IT

This paper is for Information Technology (IT) professionals working in companies and agencies that are considering the purchase of DataWalk software. The paper is intended to provide an introduction to DataWalk and address key topics of interest to IT executives, architects, and infrastructure/operations managers during their product evaluation process. The paper is organized around the basic IT functions of Planning & Governance, Delivery, and Infrastructure & Operations.

www.datawalk.com

Executive Summary

DataWalk is an enterprise-class software application and platform, for quickly integrating and analyzing vast amounts of data spread across your various data sources. DataWalk enables both business and IT organizations to improve their efficiency, productivity, and competitiveness, while also meeting key technical and security requirements.

DataWalk meets a variety of needs across your organization by enabling you to fuse, clean, normalize, and connect your vast amounts of siloed data into a single source that can be used to feed AI/ML applications, Business Intelligence tools, and various other applications. DataWalk also enables you to perform sophisticated analyses of your data via complex no-code queries, graph algorithms, link analysis, machine learning, and more. DataWalk is also:

- **Cost Effective.** DataWalk is typically a fraction of the cost of comparable alternatives, and some customers have been able to fully recoup their DataWalk investment in less than one month.
- **Secure.** Our software has been successfully deployed in highly sensitive environments including national intelligence and investigative agencies; the U.S. Departments of Defense, State, and Homeland Security; and leading banks.
- **Easily Integrated.** DataWalk seamlessly fits into your environment and easily integrates with your other data sources and applications.
- **No-code.** You can create and run complex queries and graph algorithms without coding, through an intuitive visual interface.
- **A Platform.** You can build applications to use DataWalk data and analyses, and have them safely run inside the DataWalk platform while leveraging DataWalk's capabilities. This can dramatically accelerate application development.
- **Easily Customized.** DataWalk can easily be customized/configured for specific types of users and/or applications.
- **Easy to Use.** DataWalk is an easy-to-use, easy-to-learn system that helps minimize the support burden for IT. When needed, DataWalk support is standing by to help 24x7.
- **Efficient.** DataWalk helps IT organizations maximize efficiency by empowering users to do their own analysis; adhering to common enterprise de-facto standards and processes; and minimizing resource requirements.
- **Highly Regarded.** DataWalk software consistently earns 5-star reviews from our customers in forums such as Gartner Peer Reviews. See what our customers say about us.

Table of Contents

What is DataWalk?

DataWalk is an enterprise-class software application and platform, for integrating and analyzing vast amounts of data spread across your various data sources. DataWalk enables you to:

- Efficiently fuse vast amounts of data across any or all of your internal data silos and external data sources
- Have a single source of clean, normalized, connected data that is reorganized around understandable business objects in a flexible ontology
- Provide instant access to that data for both users (through a UI) and applications (via API)
- Quickly configure and execute no-code complex queries, scores, graph algorithms, and link analysis to find hidden patterns, trends, networks, and connections across your vast amounts of complex data

Unique Capabilities

DataWalk utilizes proprietary technology, with over ten patents in various stages of approval. DataWalk's key capabilities include:

- **Data Integration.** DataWalk's graph technology is inherently a great fit for integration with other systems and data sources.
- **Knowledge Graph.** Data from your internal data silos and external sources is fused and reorganized around understandable business objects on a knowledge graph called the Universe Viewer. Here data can be integrated, connected, modeled, and queried.
- **Flexible Logical Data Model.** The DataWalk data model (visualized via the knowledge graph) can easily be modified, without impacting system operations.
- **Complex Querying.** Unlike traditional SQL databases, which cannot execute complex queries with many joins (e.g., recursive queries), DataWalk's unique technology supports linear performance through more than 600 joins. For details on DataWalk's query performance, see the paper *DataWalk Performance Test Results: Deep Queries*.
- **Flexible Scoring.** Leveraging DataWalk's capabilities for complex querying, it's easy to create and execute highly flexible scores (e.g., for risk scoring, etc.)
- **Graph Algorithms.** DataWalk has demonstrated superior performance compared to leading graph databases for graph algorithms such as clustering (which automatically identifies patterns that may indicate things like organized crime), find-paths (which automatically identifies how distant objects are connected), and others. DataWalk is uniquely suited for quickly executing graph algorithms across vast amounts of data without coding. For details, see the paper *DataWalk vs. TigerGraph Performance Benchmark: Find Paths Algorithm.*
- **Big Data Handling.** DataWalk is architected to efficiently handle billions of data records and many terabytes of structured and unstructured data.

A Must-Have for The Business

DataWalk delivers compelling advantages to the business:

- As a data platform for AI/ML and other applications, DataWalk enables the generation of superior results by integrating your siloed data into a single source of high-quality input data that is clean, normalized, and connected, with matching entities resolved.
- For anti-fraud applications, DataWalk enables remarkably accurate detection of fraud, identification, and monitoring of organized crime groups, and powerful capabilities for conducting complex investigations. The savings from avoiding payments to fraudsters has enabled customers to recoup their DataWalk investment in as little as one month.
- For anti-money laundering (AML) and Know Your Customer (KYC) applications, DataWalk can supplement your existing systems and enable you to quickly implement difficult new scenarios and respond to urgent regulatory demands, as well as accelerate complex investigations. DataWalk customers have saved millions of dollars by avoiding regulatory fines.
- For intelligence analysis applications in both private and public-sector organizations, DataWalk provides breakthrough new capabilities to fuse desired internal/external data sources and easily obtain a comprehensive view of people, events, or anything else. This can directly enable taking action to avoid potential threats and/or reduce criminal activity.

Proven In The Marketplace

DataWalk has been successfully deployed in demanding environments with government and commercial customers on five continents, including leading banks, various U.S. government agencies, and many other high-profile organizations.

DataWalk has been well received by our customers. A number of them have published reviews on forums such as Gartner Peer Insights, confirming the value, capability, and reliability of DataWalk software. And top-25 U.S. bank Ally Financial recently presented us with their Technology Disruptor award.

Architecture

The diagram on the next page illustrates the architecture of the DataWalk platform.

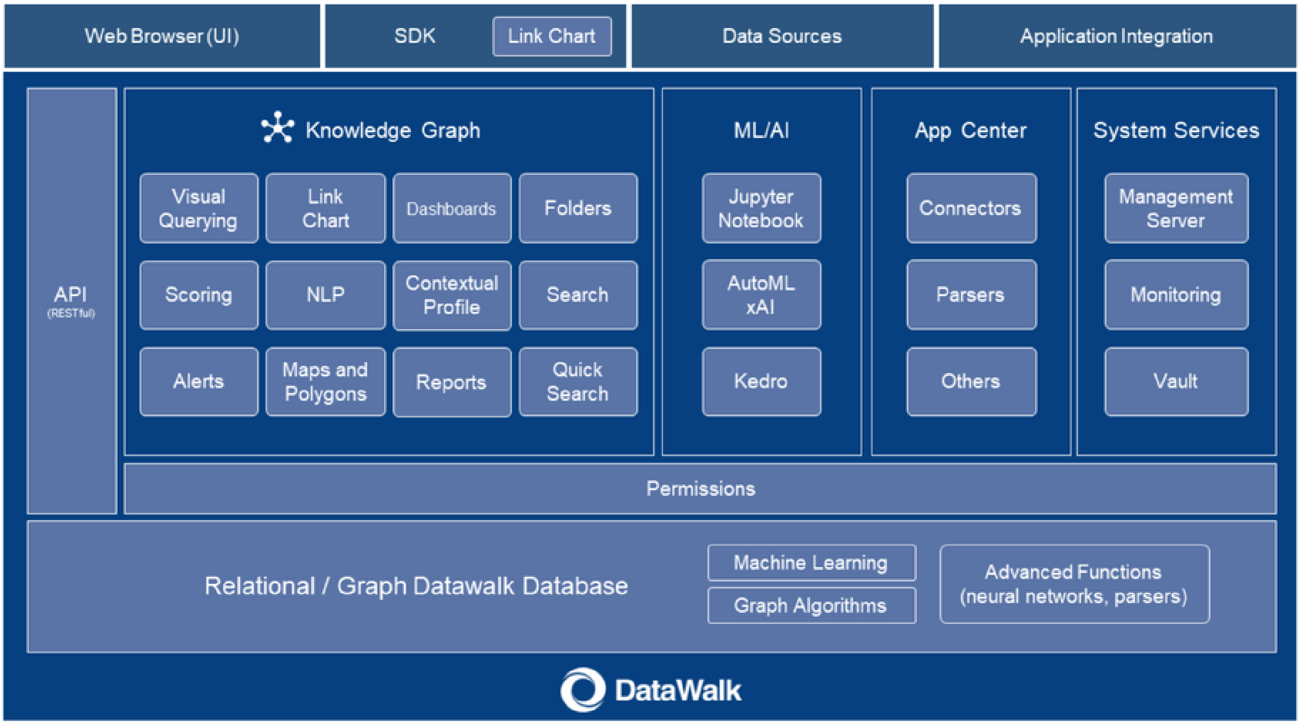

Figure 1. DataWalk architecture diagram

Users and administrators access DataWalk via a browser-based interface, such that no client-side software is required.

The heart of the system is the DataWalk Universe Viewer, which is a knowledge graph where data is integrated, modeled, visualized, and queried. In the Universe Viewer data can be restructured around understandable business objects (e.g., people, customers, transactions, and anything else), where data types are represented as nodes and data relationships (connections) are represented as edges.

The DataWalk App Center is a DataWalk facility — effectively an application framework — that enables plugins/programs ("apps") to run safely inside the DataWalk application/platform. These apps can be used to import/export data from/to other data sources/systems, extend system functionality, create your own custom features, enrich data, and perform various other functions. In addition to the apps already supplied by DataWalk, your team (with the required certification) can generate its own apps in Python.

DataWalk's AI/ML facility delivers machine-language capabilities often required by data scientists. It provides access to embedded AI/ML functions to be applied onto the Knowledge Graph. Data available via the Knowledge Graph can be extended by ML features that are calculated based on attributes or dependencies (realized by virtual columns). DataWalk also embeds Jupyter Notebook and Kedro as development and runtime environments for data scientists.

Many of the unique capabilities of DataWalk are enabled by the DataWalk database, which is an embedded, self-maintaining, scale-out, hybrid graph/relational database. Performance and/or capacity can be increased simply by adding additional nodes to a DataWalk cluster.

For Planning and Governance Teams

Supporting the needs of your IT Planning and Governance function, DataWalk:

- Can be applied to a variety of use cases across your organization
- Helps maximize IT efficiency by enabling self-service data access and graph analysis, and by simplifying the data preparation process for machine learning
- Adheres to key technology standards and de-facto standards
- Is highly cost-effective compared to comparable alternatives, with customers recouping their investment in as little as two weeks
- Is an easy-to-use software application with minimal training and support requirements
- Supports high availability and disaster recovery requirements
- Is a ready-to-deploy platform and application, with far faster time-to-solution than attempting to build your own
- Is a nimble company that is easy to do business with

Fits For A Variety of Use Cases

There are multiple usage modes of DataWalk, and the system can support a variety of use cases across your organization:

- **An application,** providing end-users with the ability to access and analyze data via an intuitive user interface for various use cases. These include:
    - Users who simply want to easily search the database and view the results via one of DataWalk's several facilities for 360-degree views.
    - Analysts who, in addition to the above, want to conduct complex querying, link analysis, and other sophisticated analytic work.
    - Data scientists who need a feature store and an environment to execute ML models with instant feedback. DataWalk embeds Jupyter Notebook and Kedro as ML development and runtime environments.
    - Data entry staff who want to enter data directly into DataWalk
    - Leaders who want to view dashboards that summarize KPIs. Dashboards can be generated directly in DataWalk, though it's more common to utilize DataWalk as the back-end data store for Business Intelligence tools such as Tableau and Qlik.
- **A platform,** enabling you to:
    - Have a single source of clean, normalized, connected data, reorganized around understandable business objects, and which can be leveraged by other applications across your enterprise. DataWalk in effect can serve as a logical database for vast amounts of data.

- Utilize DataWalk as part of an enterprise workflow. With the DataWalk API and other facilities you can pass data — and results of various DataWalk analytical facilities — between DataWalk and other systems. DataWalk can trigger events to external systems and share results of automated analyses.
- Build and/or integrate Python applications that safely run inside of DataWalk and can use data from the system.
- Create and deploy web apps within DataWalk that can utilize data from the system and enable new data interfaces. These web apps are fully deployed and managed by the DataWalk platform (e.g., auth, logs, host, HA, SSL, deployment)
- Use DataWalk as a back-end data store for live querying from Business Intelligence (BI) tools (e.g., Tableau, Qlik) to generate dashboards.

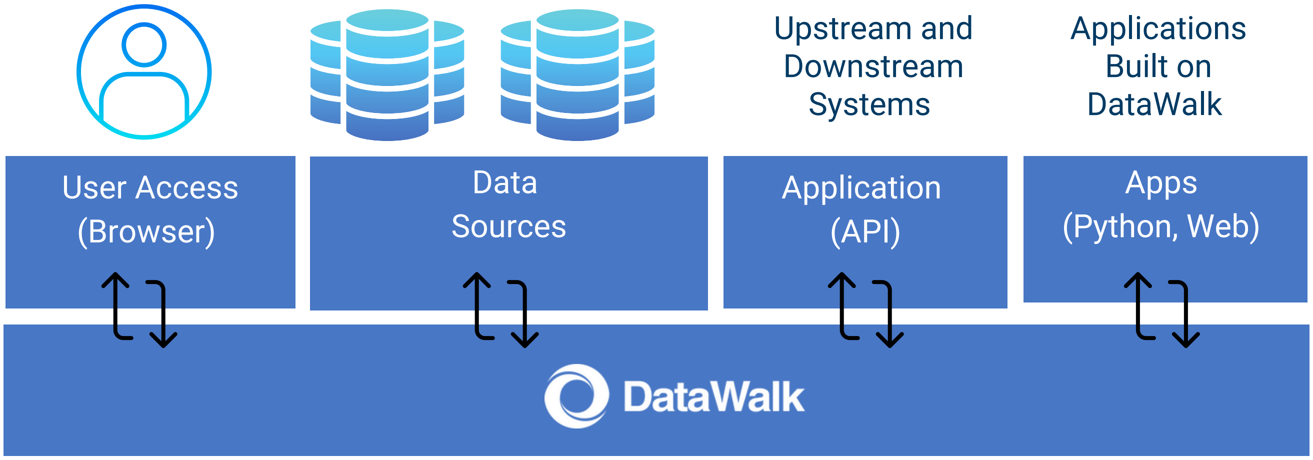

Figure 2. DataWalk is both an application and a platform.

A Logical Database

In addition to being an application that end users can utilize via an intuitive user interface, DataWalk can also be considered to be a logical database (specifically a graph/relational hybrid). It can handle complex queries from the relational world, as well as highly recursive graph queries, without copying data.

DataWalk can be queried in either of two ways: via the UI, where no coding is required (i.e., there is no need to learn a new querying language), or via API. The use of DataWalk requires no database expertise nor knowledge of any query language (such as SQL or Cypher) and the DataWalk database requires no tuning or optimization.

DataWalk supports both structured and unstructured data, and includes various facilities to support effective text processing and analysis.

Application Development Environment

If desired, you can build applications that run inside DataWalk, specifically for using data and/or analytical products that are generated in DataWalk. You can develop such applications in any language, and encapsulate them in Python.

Application development capabilities are available to programmers who have completed the DataWalk App Center Expert certification.

Developing applications to run in DataWalk dramatically simplifies and accelerates application development, you can leverage the various DataWalk facilities for security, high availability, etc., instead of building them yourself. This is illustrated in Figure 3 below.

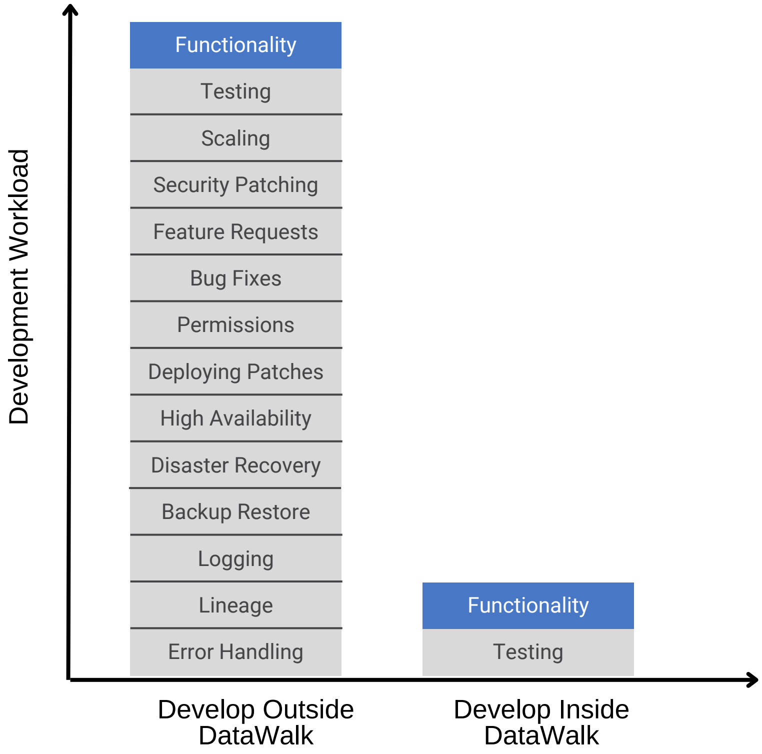

Figure 3. Development workload dramatically reduced with DataWalk.

Architectural Fit and Standards

DataWalk can operate as a stand-alone system, though it is designed to be a component of an enterprise workflow.

As such, DataWalk can serve as a data platform that provides you with a single source of clean, normalized, connected data representing any or all of the data in your organization. This enables

DataWalk to serve as the data source for ML/AI applications, Business Intelligence applications (e.g., Tableau, Qlik, etc.), and others. DataWalk is particularly valuable in environments where understanding connections between data is important.

In addition, DataWalk provides advanced capabilities for identifying and analyzing patterns and connections across all your data.

If you have existing data warehouses or a data lake, DataWalk can effectively leverage and coexist with these facilities. If you have neither, then DataWalk can serve as an alternative to either.

DataWalk software runs on commodity Linux servers, either on-premise or in the cloud (e.g., AWS, Microsoft Azure, Google Cloud Platform). It utilizes cloud-compatible storage, either on-premise (e.g., NetApp StorageGRID, Pure Storage FlashBlade, Dell ECS, or Vast) or in the cloud (as above). DataWalk is targeted to be containerized in Q1 2024, and a SaaS version of DataWalk is on our product roadmap.

Easily Integrated With Other Sources and Systems

DataWalk is architected to be a component of an enterprise workflow, and can integrate with your existing systems and workflows via a published API.

DataWalk's hybrid graph/relational technology is ideal for data integration. DataWalk has a variety of existing connectors available, and for other systems with a structured interface, DataWalk can typically connect to a new system within just a couple of days of work by a DataWalk field engineer.

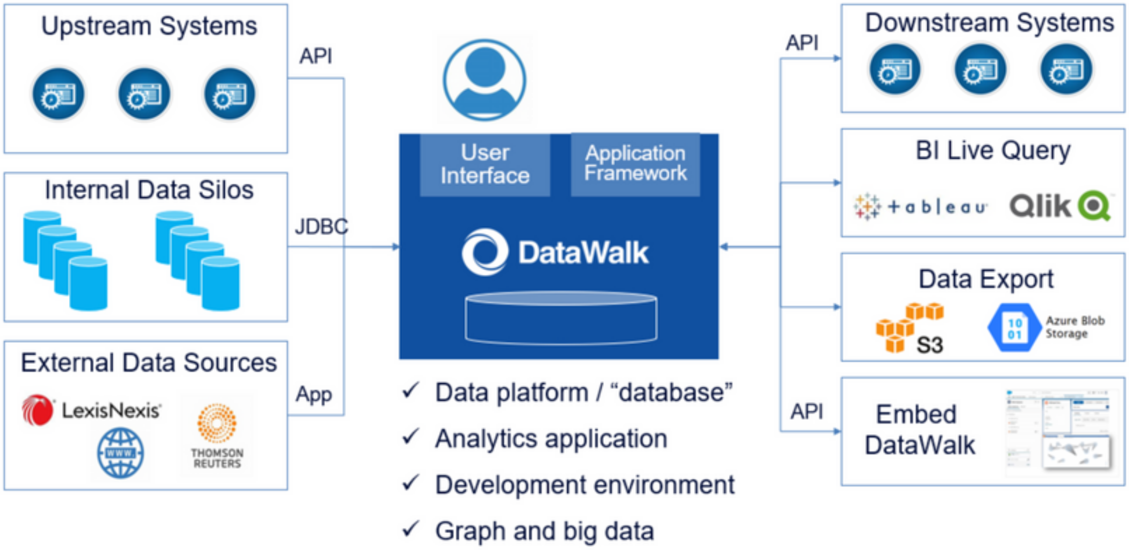

Figure 4. DataWalk as a component of enterprise workflows.

Adheres to Standards and De-Facto Standards

DataWalk:

- Runs on standard Linux-based servers (including SELinux enforced)
- Can be deployed on popular cloud services (e.g., AWS, Azure, Google Cloud Platform)
- Can utilize cloud-based storage (e.g., AWS S3)
- Supports SAML, LDAP, Kerberos, and X.509 authentication options
- Supports FIPS, CGIS, DISA STIGs, and other security standards

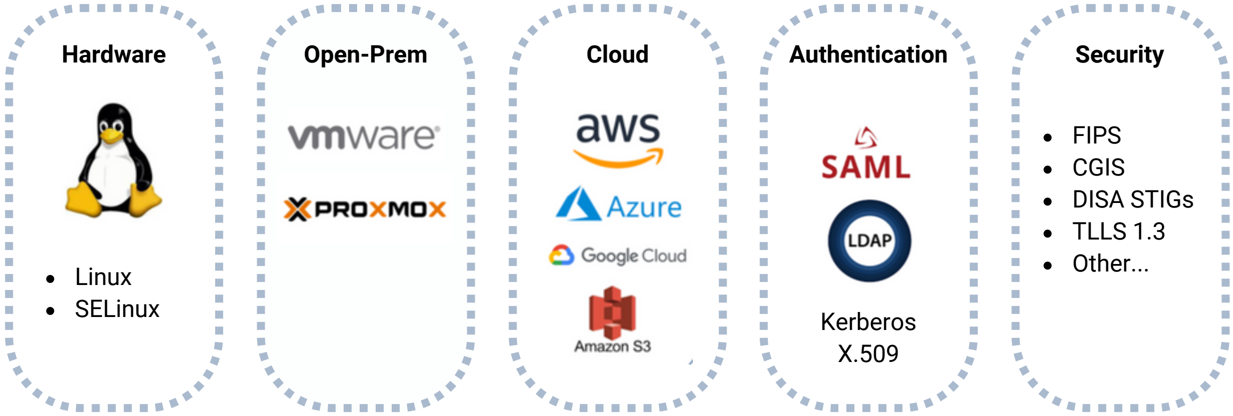

Figure 5. DataWalk support of standards and de-facto standards.

Going beyond adhering to just minimal standards, DataWalk has demonstrated full compatibility under a wide range of infrastructure setups, greatly facilitating IT's ability to meet mandated corporate technology standards and policies.

Helps Maximize IT Efficiency

As an application, DataWalk is a self-service analytics platform. End users can directly create/run queries and graph algorithms; create 360º-views of any data element; create alerts; perform link analysis; and execute various other analytic functions themselves. This offloads IT from having to perform such work on behalf of end users.

For machine learning applications, DataWalk simplifies the data preparation process, which is a common bottleneck.

DataWalk utilizes a web-based user interface, such that there is no client-side software to manage.

Cost-Effective With Tremendous ROI

DataWalk is a cost-effective system, typically with an acquisition cost that is a fraction of comparable alternatives.

DataWalk enables you to both address key analytic needs of the business, as well as provide a data platform for the business, all in one reasonably priced platform. Some customers have recouped their DataWalk investment in as little as one month.

Unlike some competitive systems, there is not a huge professional services component associated with DataWalk. It is a commercial-off-the-shelf (COTS) platform such that no custom coding is required, and new releases are delivered roughly once per quarter to provide new functionality and innovations to all customers who are on an active maintenance contract.

DataWalk is a ready-to-go platform and provides superior economics relative to the option of attempting to build a similar system yourself. With DataWalk you can start getting results and generating return on your investment, instead of waiting on the multi-year development effort of a custom application that may deliver barely a subset of DataWalk's functionality.

Infrastructure costs are minimized as DataWalk runs on commodity Linux servers, and typically requires far less CPU/RAM resources than alternative systems. For details, contact your DataWalk representative.

Easy To Use

DataWalk's ease of use helps to minimize IT's resource requirements and support burden.

The software is intuitive for end users, with a visual "no code" interface for querying, entity resolution, and other functions. User training of base DataWalk capabilities typically requires only a few hours to complete (via DataWalk's e-learning course), plus another few hours specifically for training users on your data for your specific use case. Users can perform self-service analytics, thus avoiding any need to rely on IT for data analysis.

DataWalk is also easy to administer, with admin functions easily configured via a graphical user interface and JSON.

Proven Data Security Model

Your data is safe in DataWalk.

DataWalk supports military-grade security, and has been deployed in data-sensitive organizations, including national intelligence and investigative agencies; the U.S. Departments of Defense, Justice, State, and Homeland Security; and leading banks.

DataWalk sits behind your firewall, leveraging your existing security measures. It also supports highly granular permissions, ensuring that users see only the data you want them to see.

For details, request the *DataWalk Security Guide* from your DataWalk representative.

Highly Reliable With HA and DR Options

DataWalk is highly reliable software. In addition, DataWalk can optionally be configured in a high availability (HA) configuration (which minimizes the risk of unplanned downtime due to a hardware failure) or a disaster recovery (DR) configuration (which minimizes the risk of extended unplanned downtime due to a site-wide outage). For details on HA and DR configurations, see the *DataWalk Sizing Guide.*

Backups and data model changes can be done without interrupting system operation.

Build or Buy?

Utilizing DataWalk software provides a highly attractive alternative compared to building a comparable solution yourself. Utilizing DataWalk provides measurable ROI to the business immediately, instead of after the long delay normally associated with attempting to build a suitable alternative.

The DataWalk team has invested hundreds of man-years developing this software, and we have registered over ten patents on it. We’ve also established rigorous processes for maintaining our software on an ongoing basis. For most organizations, the process of developing comparably sophisticated graph analytic capabilities (on top of a graph database) for enterprise-class operations, such as data integration and big data analytics, would take years, carry a very high risk of failure, and otherwise be highly impractical.

For more on this topic, see the papers:
Graph Analytics Applications - Build or Buy?
Investigative Analytics Software: Build Or Buy?

For Delivery Teams

Enabling your IT Delivery team to provide an effective solution aligned with business needs in a timely and efficient manner, DataWalk:

- Is off-the-shelf software that enables you to avoid the delays and challenges associated with custom development
- Requires minimal IT resources during deployment

- Demonstrates remarkable flexibility to meet your specific business needs, even if they change during or after implementation
- Is designed to easily integrate with new data sources and systems. Typically new integrations can be completed within a few days — or even just a few hours — of work by a DataWalk engineer.
- Utilizes a proven methodology for system implementation, ensuring that the solution meets your requirements, is deployed as quickly as possible, and will be successful on an ongoing basis
- Includes data management facilities, such as built-in transformations (e.g., "ELT" instead of "ETL"), normalizations, and data modeling designed to support vast amounts of complex data

COTS Software

DataWalk is commercial-off-the-shelf (COTS) software. As such, its product enhancements are delivered via regular software releases (roughly every two or three months), and made available to all customers on a maintenance contract. With this approach, you are provided with a constant stream of innovation, without the delays, complexities, and maintenance challenges normally associated with custom software.

Easily Customized

DataWalk is a remarkably flexible platform that can easily be customized to meet your specific business needs and workflows. It offers you a number of options that enable you to configure the system as needed for different users and use cases. Its flexible logical data model technology enables you to modify the data model itself. You can also enable or disable features for specific users, and enable users to have different views of the same data.

Minimum IT Resources Required to Deploy DataWalk

Deploying DataWalk is a fairly standard process, consistent with typical requirements of enterprise IT administration. To ensure that the deployment is successful and positions your organization for ongoing success with the platform, it's strongly advised that you have the following resources in place:

- A **lead administrator** (a fraction of a typical FTE resource) with the bandwidth allocated to:
  - Be trained at the start of the project
  - Participate throughout the deployment on a low-intensity basis
  - Serve as the lead system administrator after deployment is completed
  - Deploy system upgrades (though much of the process is automated)
- A **resource** to support deployment of DataWalk Management Server operations
- A **data engineer** to provide access to your data sources when needed
- A **security team** engaged for data connectivity and penetration tests (if needed)

Note that DataWalk is a self-tuning system, so there is no need for the types of database tuning and maintenance activities that are common with other systems.

Administration Tasks and Skills Required
Following are the specific tasks and skills required to administer the DataWalk system during the deployment stage.

- Administrator tasks:
  - Installation and configuration
  - System updates and monitoring system performance
  - Storage system management
  - Automating tasks and scheduling jobs
  - Configuring and troubleshooting network infrastructure
  - Security monitoring and auditing
  - Server configuration
  - Working with virtualization and containerization platforms to create and manage virtual machines and containers
  - Troubleshooting and debugging
  - Monitoring and performance tuning
  - Automation and configuration management
- Administration skills required:
  - Strong knowledge of Linux administration
  - Proficiency in command-line interface (CLI) usage and shell scripting
  - Understanding of networking protocols and services
  - Knowledge of security best practices and system hardening techniques
  - Familiarity with virtualization and containerization technologies
  - Ability to troubleshoot and debug system issues

Deployment Methodology

DataWalk has developed a project implementation methodology that has been proven in demanding environments with leading organizations in both the public and private sectors. The DataWalk approach is highly flexible and based on agile methodologies. Our approach - shown below - minimizes deployment time (typically 2-20 weeks), ensures that the solution addresses your key needs, and ensures that your organization is positioned for success with DataWalk after the deployment process is completed. For details, ask your DataWalk representative for the *DataWalk Project Implementation Guide.*

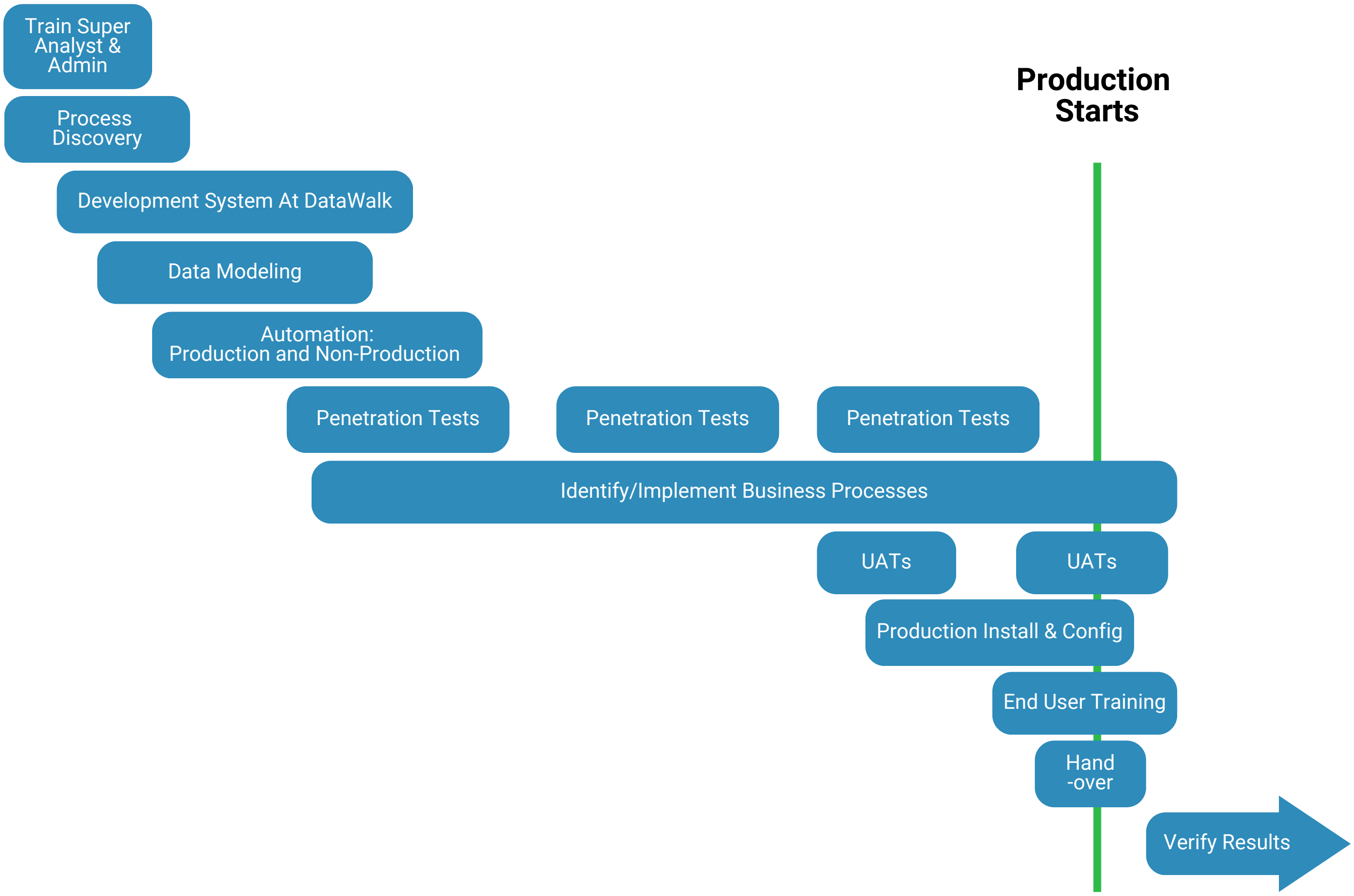
Train Super Analyst & Admin
Process Discovery
Development System At DataWalk
Data Modeling
Automation: Production and Non-Production
Penetration Tests
Penetration Tests
Penetration Tests
Identify/Implement Business Processes
UATs
UATs
Production Install & Config
End User Training
Hand -over
Verify Results
Production Starts

Figure 6. The DataWalk deployment methodology.

Extreme Flexibility to Adapt To Changing Requirements

It is not uncommon for solution requirements to change during the deployment process, as well as after implementation. DataWalk is a highly flexible software solution ideal for quickly adapting to such changing requirements. The data model, risk scores, permissions settings, feature access, saved queries, and various other system facilities and analytical products can all be easily modified by your team, without requiring coding or professional services.

Scalable To Many Billions of Records

DataWalk is based on a scale-out architecture, such that capacity and performance can be increased simply by adding additional nodes to a DataWalk cluster. DataWalk has published performance test results that reflect fast query performance for 495 billion objects (13TB of structured data) and linear performance, scaling as nodes are added. For details, see the DataWalk papers
DataWalk Performance Test: Exceptional Results With Vast Amounts Of Data
DataWalk Performance Test: Multi-Node Scaling.

Data Management

ELT

Dirty data is a common challenge, and DataWalk helps address this issue by offering the option to transform your data after it is imported into the system. This enables data to be connected and analyzed, without requiring data owners to clean their data first. Effectively this is an “Extract Load Transform” (ELT) approach, which can significantly simplify and accelerate your data preparation workflows.

If you currently utilize an ETL solution to transform your data, then DataWalk can easily incorporate your already-transformed data.

Data Ingestion

DataWalk uses a flexible, adaptable, and generic method for ingesting content. To retrieve information from a relational database you can utilize JDBC. Via the App Center there are a number of available connectors to various data services and commercial solutions. You also can easily configure connectors based on JSON or XML-based REST interfaces. External systems can register ingestion events in DataWalk, after which DataWalk will reach into the registered source when appropriate.

DataWalk can also easily ingest data residing in data lakes and flat files (e.g., Parquet, CSV, Avro, etc.).

You can ingest any desired data with DataWalk, though the recommended best practice is to ingest only the specific records and attributes that you require for analysis. DataWalk keeps track of all changes and automatically adjusts indices to ensure that data is consistent.

Ingesting data in DataWalk enables dramatically higher performance for complex queries, graph algorithms, scores, and other complex analytical functions.

Data Normalization

DataWalk includes apps (in the App Center) for addressing common data normalization challenges, such as normalizing addresses and phone numbers, and transliteration. These capabilities enable you to generate a single data set (i.e., list) of customers, phone numbers, identification numbers, or anything else that spans the entire enterprise—without coding. This in itself can represent a significant contribution to the business.

Data Modelling

The data model is represented on the Universe Viewer knowledge graph, where data sets and their interconnections can easily be created and managed. DataWalk offers a flexible logical data model, where changes can be made to the model without disrupting user/system operations. Modeling changes can be done both programmatically and via the UI.

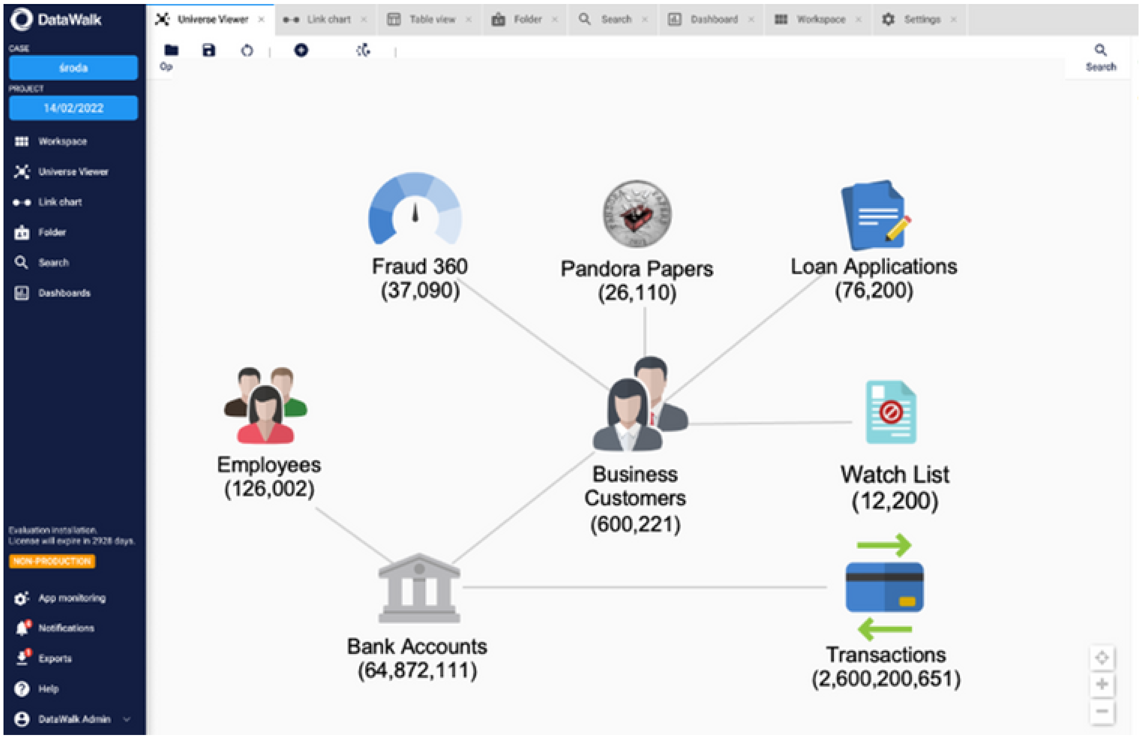

Figure 7. Example of the DataWalk Universe Viewer

Entity Resolution

DataWalk includes an entity resolution facility that enables you to easily identify and merge possible matching records using advanced fuzzy matching algorithms and a variety of techniques via a highly flexible rules engine. This facility operates at scale, with the ability to perform entity resolution across billions of records. If desired, matching records can be merged either manually or automatically.

For Infrastructure and Operations Teams

Enabling your IT Infrastructure and Operations teams to ensure stable operation of the solution, DataWalk:

- Minimizes the risk of issues when changes are made to the production system, via robust promotion pathways processes that can easily be adjusted to your data governance rules
- Requires minimal IT resources to support and maintain DataWalk
- Can be deployed either on-premise or on-cloud, using Linux servers and cloud-enabled storage
- Provides online backups to help maximize data availability
- Is designed to seamlessly integrate into enterprise environments

Virtualization and Container Management

DataWalk can work in a virtualized environment. Installation can be performed with any of the following virtualization providers: VMWare, KVM (Proxmox), Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform (GCP).

The DataWalk Management Server component can provision required VMs itself or rely on VMs that are prepared upfront. As part of the installation, DataWalk-specific containers are deployed on those machines. DataWalk itself is not yet containerized, though this is planned for delivery by Q2 2024.

Easily Upgraded

System upgrades can easily be deployed via the DataWalk Management Server, via processes that are designed to minimize the risk of issues when changes are made to the production system.

The DataWalk Support team provides comprehensive 24x7 support to help quickly resolve any questions or issues that you may encounter with our software.

DataWalk software provides your IT team with full control of the environment, including the flexibility to change system configuration, add new data sources, and easily plug applications on top of, or in, the platform.

Minimum IT Resources Required to Support and Maintain DataWalk

After deployment, DataWalk typically requires minimal incremental IT resources:

- A **system administrator** (small fraction of an FTE) to conduct ongoing maintenance, as ongoing support needs are limited. This would not go beyond the typical requirements of enterprise IT administration. This same individual may be required occasionally to configure new data sources and/or make changes to system configuration (to maintain resource continuity, it would be prudent to have at least two people trained in configuration)
- When changes are to be made to the production environment, a few hours of time may be required by a **technical resource** to move those changes through the DevOps process.

Note that backups, which should be done regularly, can be scheduled to run automatically. Also note that the internal DataWalk database is self-managing, requiring no tuning or optimization.

Administration Tasks and Skills Required

Following are the specific tasks and skills required to administer the DataWalk system after deployment, during ongoing support and maintenance.

- Administrator tasks:
    - Production:
        - Monitor available resources such as CPU, disk, and network to ensure they meet production requirements.

        - Ensure that backups are regularly created and stored.
        - Allocate space for these backups outside of the current infrastructure.
        - Keep up-to-date with security updates for the operating system.
        - Monitor data flow, if applicable.
        - Review and document any changes to disaster recovery processes.
        - For cloud deployments, ensure that automation is functioning correctly.
            - In cloud deployments, it's typical to terminate the environment at the end of every cycle (4 months, 90 days), depending on data governance. If the restore function fails, this could be problematic.
        - If there is automation for User Acceptance Tests (UATs), ensure it's current and operational.
    - Non-production:
    - Ensure there are processes in place to create non-production environments and that they are functioning correctly.
    - Adhere to the rules for promoting changes (whatever they may be).
    - Test any changes before they are introduced to the production environment
- Administration skills:

- Strong GitOps and DataWalk skills (highly recommended)
- Kubernetes administration skills (starting with DataWalk version 5.x)
- Linux administration skills (required, though applied far less than during implementation)

Given these post-implementation requirements, organizations will need:

- IT personnel to monitor resources and the technology stack
- “Super analysts” with access to non-production Management Server operations to create, destroy, promote changes, tech new users, develop new ideas, etc.

Deployment Environment

DataWalk is deployed on commodity Linux servers, either on-premise or on-cloud. On-premise options include with Internet access or without (i.e., “air-gapped environments”). On-cloud options include Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. Hybrid deployment models are also possible (e.g., on-prem servers with cloud storage).

DataWalk is not currently available in a SaaS model, though this is on our product roadmap.

DataWalk deployments include Application servers, App Center (Integration) servers, and Compute servers, each of which can optionally be configured in a clustered configuration to support high availability and/or increased capacity. In addition, a Management Server is required, and is further discussed below.

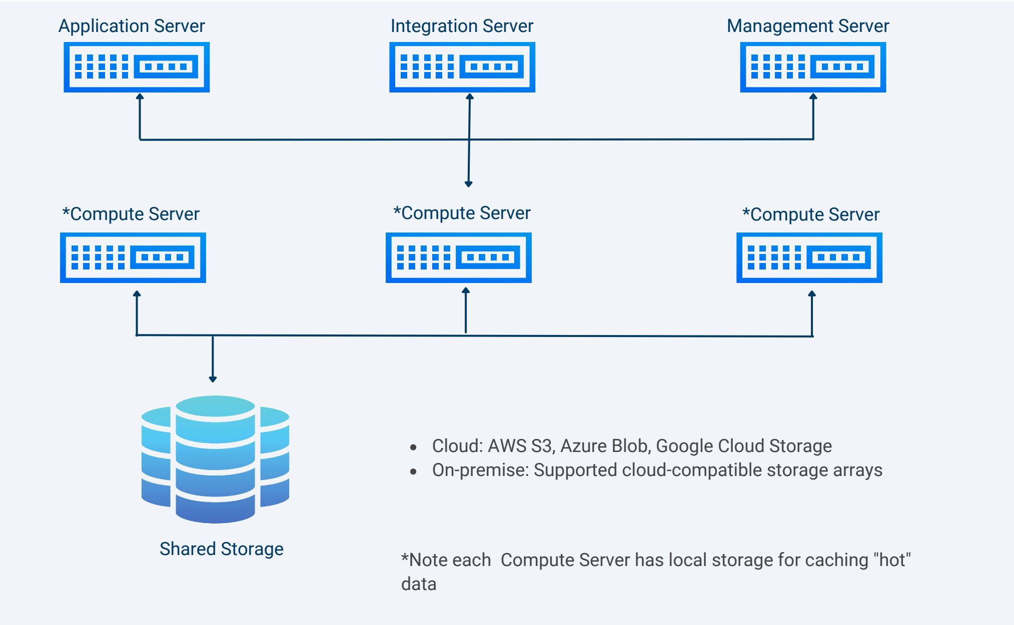
Application Server
Integration Server
Management Server
*Compute Server
*Compute Server
*Compute Server
Shared Storage
• Cloud: AWS S3, Azure Blob, Google Cloud Storage
• On-premise: Supported cloud-compatible storage arrays
*Note each Compute Server has local storage for caching "hot" data

Figure 8. Example DataWalk system configuration

Server Requirements: Commodity Linux Servers

DataWalk runs on commodity Linux servers, specifically Red Hat 8.5 or Amazon Linux 2. All servers should have processor clock speeds of at least 2.6GHz, with 3.0GHz recommended.

The minimum DataWalk production deployment (and minimum server configuration) is typically:

- One Application Server (4 cores, 16GB RAM)
- One Compute Server (8 cores, 96GB RAM)
- One Integration Server (4 cores, 16GB RAM)

Additional capacity can be added by expanding server configurations and/or by adding more application servers and/or compute servers in a DataWalk cluster. For detailed information on system sizing, see the *DataWalk Sizing Guide.*

Storage Requirements: Cloud-Compatible Storage

It is recommended that the DataWalk database reside on shared cloud-compatible storage, which may be a local storage area network (SAN) or a cloud-based storage service.

Cloud storage options include Amazon Web Services (AWS) S3, Microsoft Azure Blob, and Google Cloud Platform (GCP). Cloud-compatible SAN storage options include NetApp StorageGRID, Pure Storage FlashBlade, Dell ECS, and Vast. For these devices, all storage is in the disk arrays, on premise. Any SAN infrastructure must fulfill 60 - 80 MBit read-write per CPU core in a RAID 10 (post RAID) configuration.

Consult with a DataWalk engineer to confirm whether your available storage solution meets the DataWalk system requirements.

Storage Requirements: Cloud-Compatible Storage

It is recommended that the DataWalk database reside on shared cloud-compatible storage, which may be a local storage area network (SAN) or a cloud-based storage service.

Cloud storage options include Amazon Web Services (AWS) S3, Microsoft Azure Blob, and Google Cloud Platform (GCP). Cloud-compatible SAN storage options include NetApp StorageGRID, Pure Storage FlashBlade, Dell ECS, and Vast. For these devices, all storage is in the disk arrays, on premise. Any SAN infrastructure must fulfill 60 - 80 MBit read-write per CPU core in a RAID 10 (post RAID) configuration.

Consult with a DataWalk engineer to confirm whether your available storage solution meets the DataWalk system requirements.

Network Requirements

In general, DataWalk requires a high-speed network connection to ensure fast and efficient data transfer between nodes in the clusters. A reliable 10 Gbps network is recommended.

Non-Production Systems

DataWalk supports (and requires) DevOps automation, including dedicated development (DEV) and testing (TEST) environments in order to minimize the risk of applying new changes in the production environment.

Management Server

DevOps is enabled via a technical component called the Management Server (GitOps), which automates some or all of the following operations:

- create infrastructure for DataWalk (provision virtual machines)
- install DataWalk
- start / stop the system
- upgrade or patch existing DataWalk installation
- backup and restore procedures (see below)
- allow fine-grained access control for these operations

Backups

DataWalk delivers embedded backup mechanisms controlled by the Management Server or via shell command. The backup process can be triggered manually, scheduled, or via a customer-specific GitOps solution. Backups are for the entire application, including application configuration, critical application binaries, and structured and unstructured data.

Both full and incremental backups can be performed. Backups can be safely performed while DataWalk is running, while a restore requires a maintenance window.

All backup data is stored in a dedicated location either on your local file system or in cloud storage. The backup itself is not encrypted, such that if encryption is required, the backups should be stored on storage that is itself encrypted. Integration with enterprise backup solutions can be achieved via backup artifacts (i.e., backup files can be picked up and processed specifically by enterprise solutions).

Direct or on-the-fly data capture from the DataWalk database or DataWalk application components is not supported. Note that only DataWalk backups enable 100% recoverability.

Authentication

DataWalk has an embedded security mechanism based on groups and permissions to metadata, features, and data itself. User authentication and assignments to groups are stored internally, but can be synchronized on the fly with external systems. To do so, DataWalk can integrate with domain controllers based on the LDAP protocol, Kerberos-based domains, or SAML 2.0-based identity providers. More information is available in the *DataWalk Security Guide*.

Monitoring

DataWalk provides an internal monitoring component. Key metrics from all components are collected and presented by web interfaces based on Grafana. It is possible to integrate monitoring with your infrastructure, but it would be your own responsibility to undertake this effort.

Logging

DataWalk stores logs in files on each system component and collects them in a centralized monitoring component. There is no direct integration with SIEM, and the responsibility for this integration would be your own.

Further Reading

For additional related information, see the following DataWalk documents which are available from your DataWalk representative:

- DataWalk System Sizing Guide
- DataWalk Performance Test: Multi-Node Scaling
- DataWalk Performance Test: Exceptional Results With Vast Amounts Of Data
- DataWalk Project Implementation Guide
- DataWalk Performance Test Results: Deep Queries
- DataWalk Security Guide
- Investigative Analytics Software: Build Or Buy?
- DataWalk vs. TigerGraph Performance Benchmark: Find Paths Algorithm
- Fraud Detection And Investigation: DataWalk Or A Graph Database?
- Planning The Development, Testing, and Production Environments For DataWalk