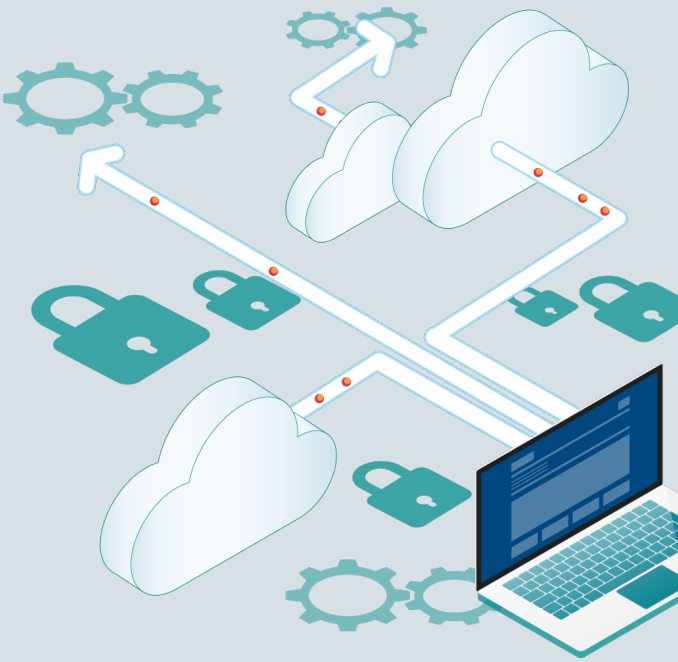
DATA Shyft

DataShyft: Maintaining Control of your Data in System-to-System Integration Pipelines



Introduction

Data is an enterprise's most valuable commodity. Protecting that data is critical to the survival of the organization. Sharing that data with partners, vendors, and contractors is fundamental to its success. DataShyft enables an enterprise to accomplish both these goals simultaneously. It allows data to be shared with partners, vendors, and contractors by orchestrating the transmission of data between systems, providing assurance that data is being used as intended by each party, and maintaining the security of that data in transit.

As an enterprise strives to improve efficiencies through data-driven decisions, it often shares data with trusted vendors, partners and customers, establishing legal agreements that define the terms of data sharing arrangements. The movement of data creates significant complexity in the enterprise's ability to manage the contractual terms and conditions, and secure and manage the data once access is provided and the data is transferred outside the enterprise.

DataShyft is a Data Integration and Orchestration Platform from Topia Technology that provides enterprises and organizations with a cross-data center, cross-cloud solution. It enables organizations to establish new business process integration pipelines that harness systems located in both cloud environments and in-house data centers for batch and stream-based data processing. To enable cross-cloud data integration and orchestration pipelines with DataShyft, an enterprise uses the DataShyft cloud service to create integration pipelines that define how data will move between their various data centers and external partners and be processed along the way. These pipeline definitions can then be

deployed onto secure DataShyft runtimes deployed in appropriate locations across the enterprise's network. Throughout the entire data integration process, DataShyft maintains the security and integrity of the data flowing through the integration pipelines.

Solution

DataShyft is a solution for orchestrating and securing data integrations between disparate systems and organizations. It leverages Topia Technology's existing tools and frameworks to enable an entirely new level of data integration, orchestration, and security. It goes beyond solutions that just move data between systems. It also provides:

- transformation and filtration of data, allowing data sharing between disparate systems,
- data streaming for near real-time transfer through the data pipelines,
- military grade encryption to protect data as it travels through the orchestration pipeline,
- better auditing of actions, data, and movement of data, and
- data governance to automatically enforce contractual terms.

DataShyft is an Integration Platform as a Service (iPaaS) hosted by Topia Technology. DataShyft's cloud platform is used to define and manage the orchestration and integration pipelines. When a customer deploys a pipeline, the pipeline definition is sent to the DataShyft Runtime container hosted on customer-controlled systems. Each runtime instantiates and executes their portion of the pipeline to achieve the desired integration and orchestration behavior. In addition to their data processing roles, the runtimes also act as the enforcement point for data governance, helping to track data flow and enforce limits on where data can go and how long it can be used. This provides the flexibility of a cloud service with the enhanced performance and security of a on-premise solution.

DataShyft comes packaged with a management suite that allows IT Administrators and policy creators to create and manage data integration pipelines, as well as view, filter, and search audit logs. The management suite offers a set of tools for creating new pipelines as well as viewing and configuring existing pipelines. Since pipeline logic can become quite complex, existing pipelines can be viewed graphically. DataShyft users can view the status of their data orchestration and integration pipelines allowing them to monitor the operation and performance of their pipelines. Additionally, the management tool will allow the enterprise to analyze and review the audit logs to understand where their data is going and verify conformance with contractual agreements.

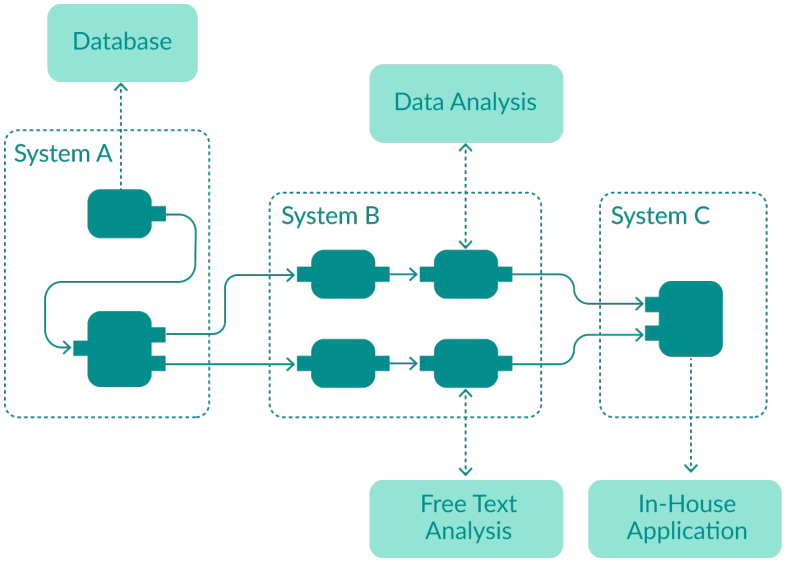
DataShyft builds on the hardened technical capabilities of Topia Technology's Kolona framework. Kolona enables organizations to create data integration pipelines that bypass the centralized brokering architectures of traditional integration solutions. Instead, it uses a disintermediated approach where

info@datashyft.com | 253.572.9712 | 1927 Dock Street, Tacoma, WA 98402

pipelines are defined from a central platform, but are deployed and run on systems participating in the data integration process. This dis-intermediated approach removes bottlenecks in the integration process and allows for more efficient use of network resources. DataShyft uses this system integration platform and toolkit to enable easy and secure integration of disparate systems.

Data Pipelines

At its core, a DataShyft Data Integration and Orchestration Pipeline is an assembly of individual components that produce, transform, filter, and consume data. The output of one component is connected to the inputs of other components. When a component returns data, that data is passed on to each component to which it is connected. Since DataShyft moves data to downstream components as soon as it is returned from upstream components, it is able to stream data through the pipeline, maintaining higher throughput and lower latency. Using this structure, data integration and orchestration pipelines of any complexity can be constructed for both one-time, batch, and streaming data flows.



In a typical pipeline, there are one or more components that read data from a source, such as a database or file service, and return that data. Those components are usually connected to transform or filter components that output new data objects based on the input data and their individual configurations. This process continues through the pipeline until data reaches components that store it into a source, such as posting to a web service or inserting it into a database. These components typically represent the end of a pipeline's data flow.

While pipelines always move data from the output of one component to the input of another, this does not preclude the creation of bi-directional pipelines. Bi-directional pipelines simply have components whose data is flowing in opposite directions across the systems, with one or both ends connected together.

info@datashyft.com | 253.572.9712 | 1927 Dock Street, Tacoma, WA 98402

Data Orchestration

Data orchestration and Integration Pipelines are assembled either from components provided by the platform or custom components created by the customer. A user selects the appropriate components and connects them together as appropriate to achieve their goal. DataShyft provides components for accessing common data storage systems and service and performing common transformation and filtering operations. Additional data access and processing components are being regularly developed and released based on customer feedback.

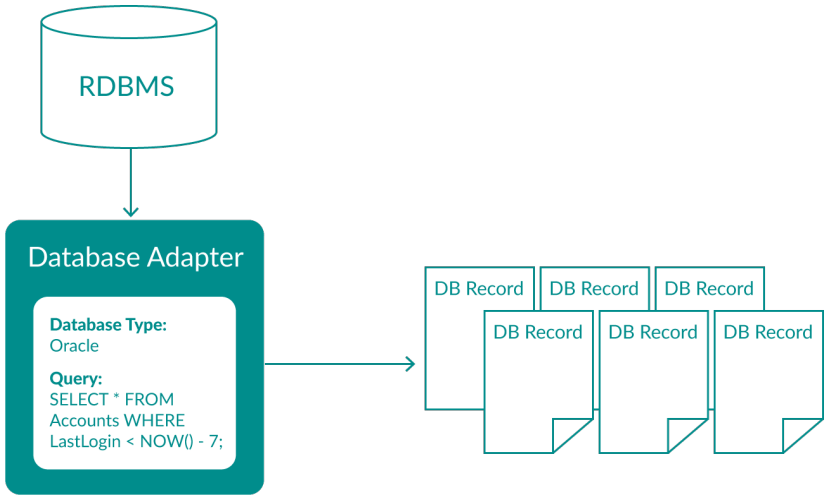
Data Governance

DataShyft includes a Data Governance System that allows for the tracking of data through an integration pipeline to maintain data provenance. By tracking data provenance, DataShyft is able to track data as it is transformed, filtered, and stored. This allows it to know what data needs to be removed when access to source data items are revoked. When access to a data item is revoked, either automatically according to programmed contractual terms, or explicitly by user or pipeline configuration, DataShyft's Data Governance components use the data provenance information to identify all data derived from the revoked items and trigger the revocation and removal of the original data as well as any derived data.

To accommodate differing scopes and requirements, the Data Governance System supports multiple back-ends for tracking data provenance. Typically, data provenance information is tracked in a enterprise-owned database for speed and security. In certain scenarios, however, a more distributed approach may be required. In these cases, data provenance information can be tracked using a Smart Contract on a blockchain, such as IBM's HyperLedger.

Data Access

The core of any data orchestration and integration system is the data access components. DataShyft provides database access components that support all the major database systems: Oracle, MySQL, MS SQL Server, etc. Any database system that has a JDBC driver can be used with DataShyft by installing the driver on the runtimes, configuring the database access components to use the



info@datashyft.com | 253.572.9712 | 1927 Dock Street, Tacoma, WA 98402

appropriate driver, and specifying appropriate queries or statements for that database. Beyond relational databases, DataShyft supports non-relational databases, such as Apache Cassandra. In addition to database access, DataShyft also provides components for accessing data stored on File Services, such as FTP servers.

Transformation

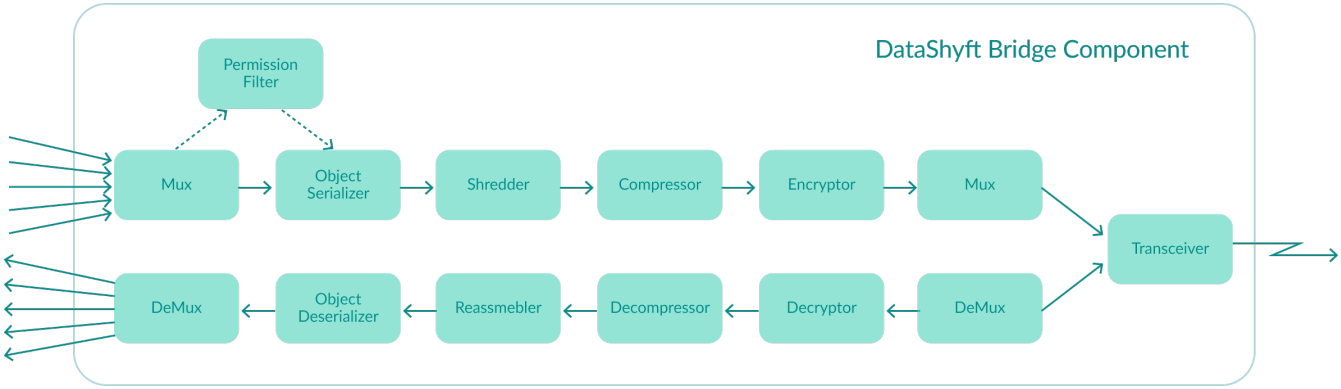
Once data has been ingested into the pipeline, data transformation is an important next step in the integration and orchestration process. DataShyft provides components that are able to transform data between standard types, such as XML and JSON, as well as components that allow custom data transformations to be defined by a pipeline user. These custom transforms are configured using a scripting language that specifies the process of transforming the input data into the desired output format.

Filtering and Masking

To protect the security of data flowing through the integration and orchestration pipelines, data filtering components are available that remove unwanted data from the data flow. Data masking components are available that can be used to mask sensitive fields within data records. The filtering and masking rules for a particular pipeline are defined by the pipeline owner. Once deployed, these rules are applied to all data passing through the components.

Data Protection

It is critical that data moving between data centers or cloud infrastructure systems be protected from unauthorized disclosure. DataShyft ensures that all data is protected using secure network bridging technology for communication between runtimes participating in a data integration and orchestration pipeline. The secure bridges use patented processes for protecting data during its traversal across the



info@datashyft.com | 253.572.9712 | 1927 Dock Street, Tacoma, WA 98402

network. The bridges negotiate a shared encryption key that is used to encrypt all traffic flowing across the bridge connection. The bridges will periodically renegotiate these encryption keys based on elapsed time or volume of data processed. This is in addition to standard protection layers such as TLS/VPNs and help to provide "defense in depth" against any attackers.

Auditing

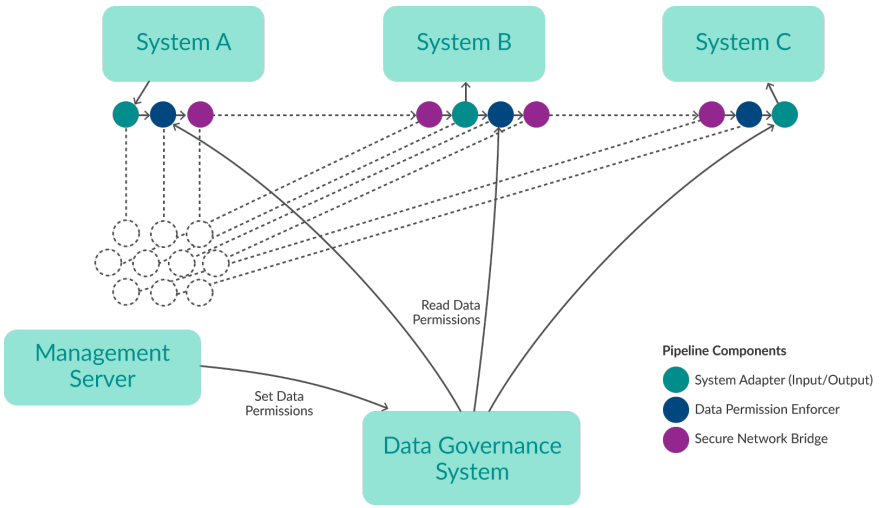
Included in DataShyft's standard component suite are auditing components that record the actions and activities that occur within the pipeline. DataShyft sends the audit logs to the DataShyft platform for storage and review. Users can review the audit logs in the future if there are questions about conformance to data security policies. An organization can specify when, where, and how activity should be recorded in the audit log by explicitly placing auditing components within a pipeline.

Example Scenarios

Business Process Automation

Enterprises have numerous processes throughout the organization that can benefit from automation. Some of these processes involve sharing data between business units within the enterprise. Some involve sharing data with external partners and vendors. Using DataShyft, the IT department works with the two outside organizations to define the appropriate data access rules and integration steps for the automation. Once the pipeline has been defined, the IT department loads the data access rules onto the Data Governance System and deploys the pipeline. This triggers the transfer of the components of the pipeline to the appropriate systems across the organizations.

Once the pipeline is deployed, it connects to the various systems needed for this integration. If configured, the data access rules are read from the Data Governance System and distributed throughout the pipeline to the appropriate components. With the pipeline fully configured, data is read from the source systems, transformed, filtered, transmitted to the destination



info@datashyft.com | 253.572.9712 | 1927 Dock Street, Tacoma, WA 98402

systems, and delivered to the data processing systems as required. Throughout this process, the pipeline provides status updates on where and how the data is flowing, and the systems to which the data has been delivered.

If the enterprise needs to change the data access rules in the future, they can update the data access rules in the Data Governance System. These updated rules will be passed to the data integration pipeline components and propagated through the pipeline.

When the data sharing agreement reaches its end, the data access rules can be updated to remove data access completely. The data integration pipeline will again detect the updated rules, read them, and reconfigure the pipeline. With the data access now having been revoked, the data is removed from the destination system storage locations where it was placed. This allows the enterprise to be confident that the data is disposed of properly.

Artificial Intelligence

Many AI-based systems require accessing large amounts of information, not all of which can be easily or cost-effectively co-located with the AI software. DataShyft provides a method for distributing the AI software and algorithms out to the systems holding the data to be analyzed. The AI modules are deployed across the network and fed with data from each of the various sources. The output, both interim and final, of these modules can be forwarded back to the primary AI system where the training results can be combined. This can result in enormous bandwidth savings since only the algorithms and model data are moved, not the entire dataset. Additionally, the latency in data retrieval for processing is greatly reduced since it no longer involves transferring the large data sets across the network for processing and analysis.

Data Privacy and Permission Enforcement

Data privacy is a topic of great concern amongst corporations, privacy advocates, and governments throughout the world. Corporations wish to use the data they have collected to improve their business processes and find new revenue streams. Privacy advocates are fighting to protect the privacy of individuals who often have little or no say in how corporations use their data. Governments are debating and passing laws codifying the privacy rights of individuals. In this environment, the ability to manage access to private customer data is becoming paramount.

Using DataShyft, corporations can define data integration and orchestration pipelines that move data between them and their partners. To ensure they conform to the privacy wishes of their customers, DataShyft allows the inclusion of filters that can mask or remove data that customers do not wish to

info@datashyft.com | 253.572.9712 | 1927 Dock Street, Tacoma, WA 98402

have shared, or which the corporation cannot share for legal reasons. Customers can then be given the option to opt-in to having their data shared, and be given the opportunity to specify which pieces of data they wish to share, and which they want to have kept private. These decisions can then be written to the blockchain as a record of the customer's wishes. The data orchestration pipeline can then uphold the customer's wishes by reading their permissions from the blockchain and enforcing those restrictions on the data flowing through the pipeline.

Because DataShyft can retrieve basic permissions from the blockchain, it is readily able to adapt to reading additional permissions from the blockchain, including from other smart contracts. This allows DataShyft pipelines to easily adapt to the changing privacy stance of individuals.

info@datashyft.com | 253.572.9712 | 1927 Dock Street, Tacoma, WA 98402