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. But sharing that data with partners, vendors, and contractors is fundamental to its success. DataShyft enables the enterprise to accomplish both of these goals simultaneously by allowing data to be shared with partners, vendors, and contractors while also 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 the enterprise strives to improve efficiencies through data-driven decisions, it often shares data with trusted vendors, partners and customers. The terms of this data sharing are established through contractual agreements. This extensive movement of data creates significant complexity in the enterprises’ ability to manage contractual terms and conditions, and secure and manage the data once access is provided and the data is transferred outside the enterprise.

Topia Technology’s DataShyft is a Data Integration and Orchestration Platform that provides enterprises and organizations with a cross-data center, cross-cloud solution. It enables organizations to create new business process integration pipelines that involve 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, the enterprise simply installs the secure DataShyft runtime on one or more systems inside their data centers or cloud environments and then uses the DataShyft cloud service to define pipelines that move and stream data between the systems inside their various on-premise and cloud data centers. 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. Instead, it provides:

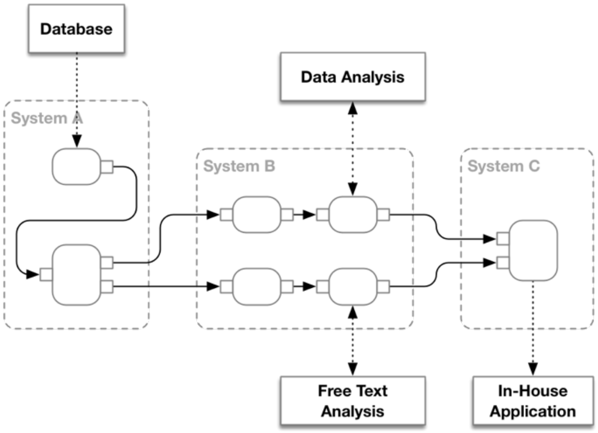
- transformation and filtration of data, allowing data sharing between disparate systems,
- data streaming for near real-time transfer through the data pipelines,
- better auditing of actions, data, and movement of data, and
- military grade encryption to protect data as it travels through the orchestration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

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 defined pipeline, the pipeline definition is sent to the DataShyft Runtime Environments hosted on the customer's own systems. The runtimes then create and execute their portion of the pipeline to realize the desired integration and orchestration behavior. This provides the flexibility of a cloud service with the enhanced performance and security of a hosted solution.

DataShyft builds on the 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 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 will use 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 and consume data and are connected together. The output of one component is connected to the input of other components. When a component returns data on one of its output channels, that data is passed on to the input channels of each component that is connected to it. 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 and maintain high throughput and low 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 on their output channels. Those output channels are typically 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 consume the data they receive and represent the end of a pipeline's data flow.

While pipelines always move data from an output channel of a component to an input channel, 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.

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 adds the necessary components to the definition and connects them together to perform the operations necessary to achieve their goal. DataShyft provides a number of components for accessing common data storage systems and service and performing common transformation and filtering operations. This includes components that allow for extracting data from and inserting data into common Relational Database Systems, such as Oracle, MySQL, and MS SQL Server. Components are provided for accessing webservices and retrieving and delivering data. Generic HTTP components are provided for those web services that don't have specific pipeline components.

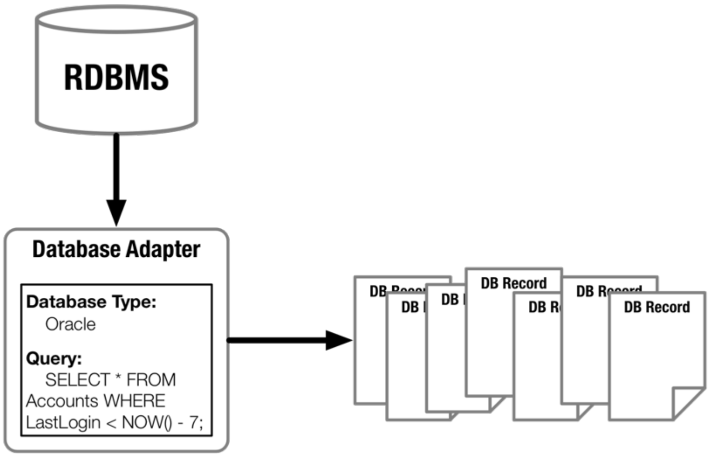
Data Security

DataShyft's Data Security components enable the pipeline to enforce contract terms and conditions at a variety of points throughout the pipeline. The Data Security Filters receive the permissions, which can be stored in a blockchain, and enforce those permissions on the data flowing through them, dropping data that downstream components and systems are not authorized to access. The Blockchain Data Security Monitor components monitor the Smart Contract for any changes to the permissions. When changes are detected, they fetch the updated permissions and send those through the pipeline to the Data Security Filters.

In addition to the data security blockchain components, DataShyft provides additional components that allow a pipeline to interact with blockchains and smart contracts in custom ways. These components are configured with the location and type of the blockchain in question as well as with any information necessary to sign transactions that will be submitted to the blockchain. When placed into a pipeline, these components can be used to either read information from the blockchain, submit arbitrary transactions to the blockchain, or invoke methods on smart contracts. DataShyft is not tied to a particular blockchain and will support the most popular blockchains used by enterprises.

Data Access

The core of any data orchestration and integration system are 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 simply configuring the database access components to use the appropriate driver and specifying appropriate queries or statements for that database. Beyond relational databases, DataShyft supports a number of non-relational databases, including Apache Cassandra, MongoDB, and Amazon DynamoDB. In addition to database access, DataShyft also provides components for accessing data stored on File Services. These components provide the ability to store and retrieve files from FTP servers, CIFS servers, and common cloud-based file repositories.



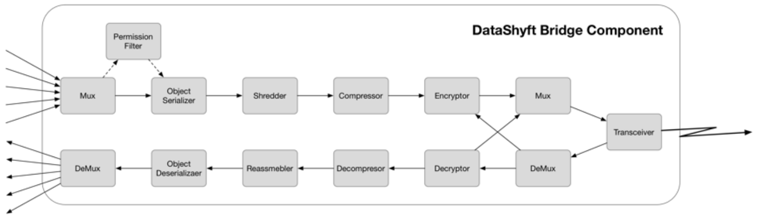
Transformation

Once data has been ingested into the pipeline, transformation is an important next step in the integration and orchestration process. DataShyft provides components that are able to transform between standard data 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. Further, 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



When moving data between data centers or cloud infrastructure systems, it is critical that the data be protected from unwanted exposure. DataShyft ensures that all data is protected by using secure network bridging technology for all communication between systems in a data integration and orchestration pipeline. These bridges use multiple layers of security to protect the information flowing over the network. The bridges use Transport Layer Security (TLS) to establish the initial secure connection between the systems, verifying certificates in both directions to validate the identity of the systems. Once the TLS connection is established, the bridges will then perform a separate Diffie-Hellman key negotiation to generate a shared encryption key that is used to further encrypt all traffic flowing across the network. As a final layer of data protection, the bridges will automatically renegotiate their encryption keys after a certain length of time or volume of data has been processed. These layers help to provide "defense in depth" against any attackers.

DataShyft users can view the status of their data orchestration and integration pipelines via the DataShyft Cloud Platform. Pipelines and their components regularly report their status back to the DataShyft Cloud platform where the data is coalesced and made available to users so they can monitor the operation and performance of their pipelines.

Enforcing Data Access Rules

Data integration and orchestration pipelines are deployed and managed by the organization that owns the data flowing through the pipeline. To enforce data access rules, the owning organization configures the integration and orchestration pipeline to include data security components in the pipeline upstream of the components that deliver data to outside organizations. If the owning organization decides to change the data permissions granted to the outside organizations, it can update the blockchain smart contract with the new permissions.

Rules can be defined regarding data retention policy on any systems that are involved in a pipeline. These rules will be automatically enforced by the Runtime, and will report the enforcement of these rules to the cloud service.

When the terms, conditions, and permissions of an agreement are updated, the changes are recorded on the blockchain inside the associated Smart Contract and propagated to the entire blockchain network. Data integration pipelines running in the network take notice of these updated access rules in their associated smart contract and adjust the behavior of their data security filters as required to ensure that these access rules are enforced. This may involve filtering the data flowing through it to protect sensitive information, or revoking access to a data source and removing shared data from systems that are no longer authorized to have it. If the permissions are changed while data is in transit, the data security components are able to receive the updated permissions and start applying them to data flowing through the pipeline. The data security component can then stop data already in transit from being delivered to an outside organization.

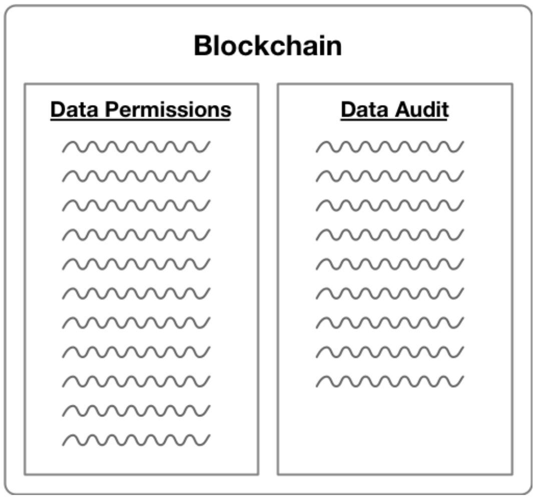
Downstream Data Governance

DataShyft changes how enterprises handle data governance. Since data integration and orchestration pipelines are able to configure the handling of data on remote systems through the runtimes on which they execute, they can be configured to provide advanced management and governance of the data in the pipeline. Basic governance includes controlling what data is passed to particular downstream systems. More advanced data governance involves the management of data that has already been delivered to a system. Such advanced data governance can include tracking the lifespan of the data on the system and automatically removing it when it is no longer needed. It can also include providing and subsequently revoking the encryption keys needed to access data to ensure that data is only available for the intended purposes. This ability to control data access across all parts of the pipeline enables a new level of data governance and reduces compliance costs for the organization.

Auditing

Included in DataShyft's standard component suite are auditing components that record the actions and activities that occur within the pipeline. DataShyft includes the ability to send the audit logs to the DataShyft platform for storage and review, or to sign the audit logs and record them within a blockchain smart contract to ensure immutability and to maintain control. In either case, users can review the audit logs in the future if there are questions about conformance to data security policies. By placing auditing components within a pipeline, an organization can specify when, where, and how activity should be recorded in the audit logs.

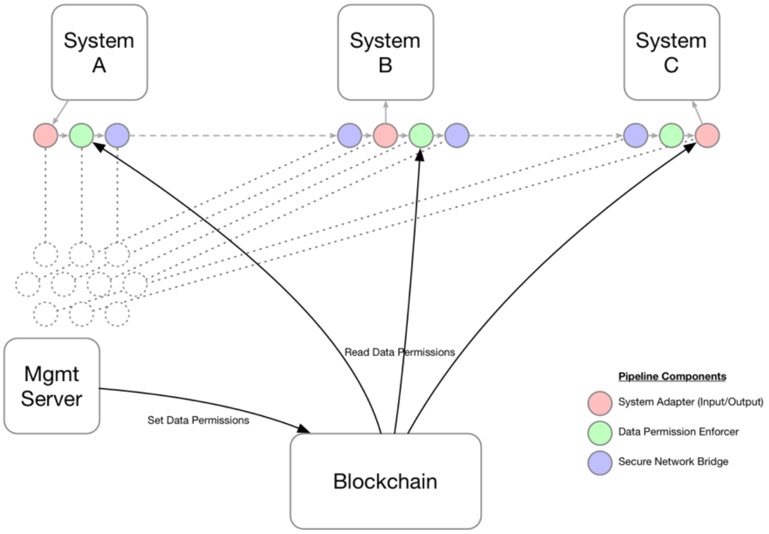
DataShyft comes packaged with a management suite that will allow IT Administrators and policy creators to create and manage data integration pipelines, as well as view, filter, and search audit logs. The management suite offers a set of tools for creating new pipelines as well as viewing and configuring existing pipelines. Since pipeline logic can become quite complex, existing pipelines will be viewable graphically. Tools will be made available to allow testing the pipelines before deploying them in a production environment. 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.



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 organizations to define the appropriate data access rules for and integration steps for the automation. Once the pipeline has been defined, the IT department loads the data access rules onto the smart contract on the blockchain, then 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 blockchain and distributed throughout the pipeline to the appropriate components. With the pipeline fully configured, data is read from the source systems, transformed and filtered, transmitted to the destination 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.

In the future, originating organization may wish to change the data access rules to limit the data received by the other. In this case, updated data access rules can be deployed to the smart contract on the blockchain. These updated rules will be detected by the data integration pipeline components, read off the blockchain, and propagated through the pipeline.

When data sharing agreement reaches its end, the data access rules can be updated to remove data access completely. The data integration pipeline will detect the updated rules, read the rules and reconfigure the pipeline, and then remove data stored on the destination system from any 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 where the data to be analyzed resides. 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 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.