

Solving the Problem of Data Silos: Process and Architecture



Run risk, compliance, and fraud detection applications on a comprehensive, global, and always up-to-date data set.

Introduction

The lack of visibility across data silos — data sources that are not integrated with enterprise systems — is a threat to business efficiency and profits in many industries. In financial services, front-office silos may develop where operations are segregated by product and region without coordination on data model design. Mergers and acquisitions may result in additional disparate silos, or regulations may require that data in one arm of the firm be inaccessible to another. When risk managers and compliance officers in financial services firms cannot see how activities in one silo are related to activities in another, the chance of rogue risk-taking, rate manipulation, or financial fraud is high. This May 2015 headline is just one example of the consequences: “Five global banks to pay \$5.7 billion in fines over rate rigging.”

Most firms have risk and crime prevention operations aimed at forestalling such headline events. But the applications they use cannot give them a clear line of sight into the data across all of the firm’s silos. Without this awareness, it is nearly impossible to recognize anomalies and make adjustments before they become larger problems. Despite the best efforts of risk managers and compliance officers, negative events still occur.

This white paper describes a data line-of-sight solution based on InterSystems Ensemble® technology. It illustrates how you can gain visibility into data and activity across all of your silos to reduce the risk of negative events.

Gaining a line of sight across data silos

As a data aggregation and conditioning service for siloed data, InterSystems Ensemble provides functionality for each step in creating a data line-of-sight solution. It includes technology for data ingestion, normalization, storage, and analytics, and provides your risk and crime detection applications with access to a comprehensive, global, and always up-to-date data set. The original data stays in its required locations within your firm.

As part of its workflow, Ensemble can take action when an anomaly is detected, such as launching a business process that automatically gathers related information and then presents that information and the anomaly to relevant personnel for a rapid response.

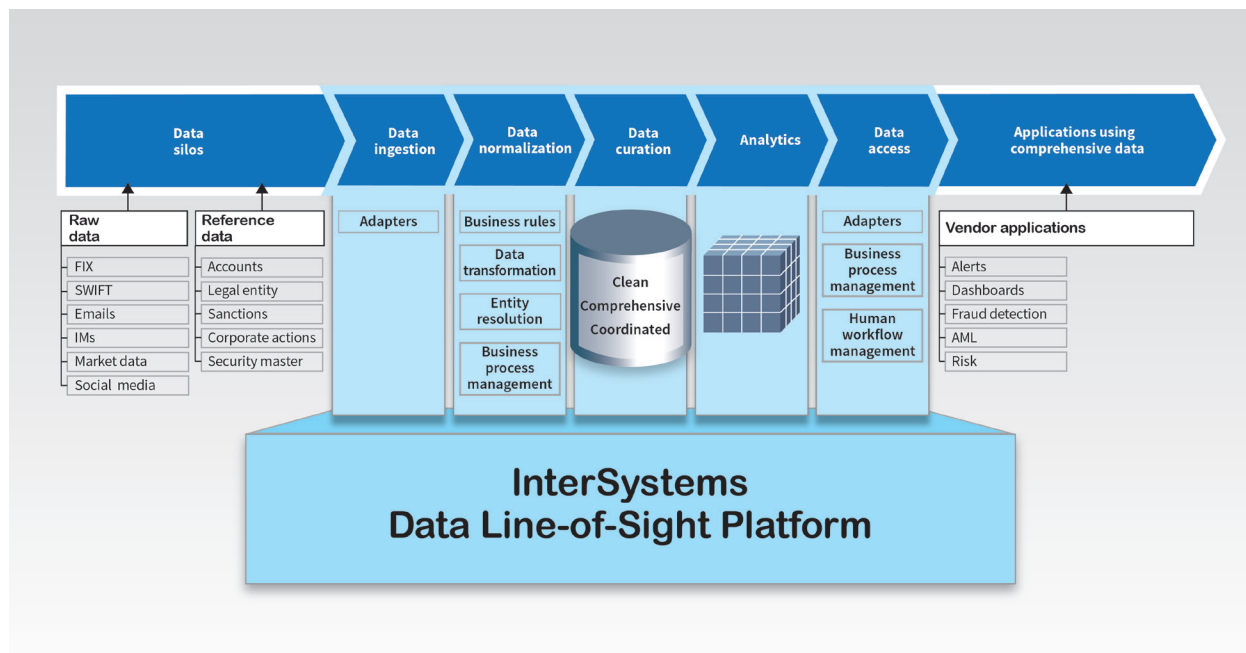


Figure 1. The arrows along the top of the illustration represent the data line-of-sight process. Data line-of-sight platform components are shaded in light blue.

Data ingestion, normalization, and curation

Ensemble ingests data from a wide variety of data feeds and silos via high-speed adapters.

For dynamic queries, such as drilling down into the details of records that remain in their original locations, Ensemble orchestrates calls to external services or sources to bring in the data.

Before aggregating data, Ensemble uses business rules and lookup tables to normalize formats and terminology. Data normalization includes entity resolution, where different values, such as variant name spellings, have been used to represent the same person or object.

All of the information ingested into Ensemble is saved in its onboard, high-performance database, along with metadata describing the information and relationships.

Analytics

As feeds that represent time-based data enter Ensemble, it is often not enough to store the data for later use, even in a normalized form. It is often necessary to analyze the data and extract value from it to maintain “state” information, such as point-in-time values and statistics. The analysis is also used for recognizing anomalies, such as when trade volume for a customer suddenly increases from its typical rate of 100 trades per hour to 1,000; or when cash transactions via SWIFT, normally \$20,000 per day, suddenly rise to \$2 million. To recognize anomalies, the system must be able to perform analytics on past data to establish benchmarks for comparison with current activity.

Not all of the important information is quantifiable, however. The data line-of-sight solution should enable analysis of textual data, such as that contained in emails, social media, and news, to extract meaning and correlate it with numeric analyses.

Access to the comprehensive and coordinated data set

To complete the line of sight, information from across the enterprise has to be made available either to existing compliance or analytics engines, directly to users of the system, or both.

The same mapping tools and adapters used to bring data into Ensemble from many sources are used to take the data from Ensemble’s internal, canonical form and feed it into these existing applications in the correct format.

Ensemble’s internal analytics engine generates dashboards based on the raw data or maintained state, and executes business rules to recognize exceptions and initiate actions or alerts.

Initiating action

Once business rules and analytics engines have identified an exception to normal activity, users can take action. Within Ensemble, a business process definition can initiate a complex chain of actions, such as gathering relevant data from several sources and generating an outbound communication in the form of a suspicious activity report that includes all appropriate information.

Ensemble business processes and workflows can incorporate human intervention where necessary to approve or validate notifications, or they may simply orchestrate calls to the external data sources.

Data line-of-sight components

Data ingestion: Adapters

The adapter technology within Ensemble enables applications and other data sources to send information to Ensemble, and for Ensemble to return information. Adapters manage the details of the relevant communications protocols and error handling, as well as maintaining and retrying connections. Ensemble adapters can be grouped broadly into three types:

- Document or application protocol adapters such as EDIFACT or SAP that provide high-level data structures that can be manipulated and routed by analysts and information specialists.
- Technology protocol adapters such as SOAP, SQL, HTTP, email, and JMS that provide developers with easy-to-use interfaces for handling input and output over these protocols, or for calling external programs without having to worry about the intricacies of the wire protocol or error-handling and retry procedures.
- Language bindings that provide programmers using Java, C++, .Net, and other languages direct access to the Ensemble data platform. While this requires more technical skills and development time, it represents the fastest possible way to load data into the data store.

Data normalization: Business rules engine

Ensemble business rules provide non-programmers with a graphical user interface for defining the evaluation and routing of requests or messages entering Ensemble. To provide a data line-of-sight, these rules validate incoming data and determine how to normalize it. Business rules also form an important part of business processes initiated when an exception is detected.

Data normalization: Transformations

Ensemble data transformations provide powerful mapping from one message or document format to another, and use lookup tables and standard functions to modify content and provide a normalized form. Ensemble applies data transformations as data is ingested and stored in the database, or dynamically transforms data as it is presented to analytics or compliance engines.

Ensemble's graphical interface for defining data transformations increases speed of development and makes these mappings accessible to non-technical analysts.

Data normalization: Entity resolution

Entity tables within Ensemble allow identities to be mapped across silos. For example, the same customer or trader identities in different silos can be mapped to one another so that data can be normalized and aggregated effectively.

Data normalization: Business process management

The business process management component makes calls to external systems, either in parallel or in sequence, to implement data workflows. These requests can be a short set of synchronous calls to query an external database or invoke a web service, or they can be complex flows of requests, each dependent on the earlier results.

Business processes can include long-running requests that take hours or days to complete. Because Ensemble stores the state of each request, business processes can survive system restarts or other interruptions.

Ensemble provides a graphical business process language to implement business process orchestration. Again, this graphical interface not only increases productivity, but also makes business process definitions accessible to non-technical analysts.

Data curation

InterSystems is a world leader in high-performance databases, including the one at the heart of the data line-of-sight solution. Recognized as a leader in Gartner's "Magic Quadrant for Operational Database Management Systems" report in October 2014, it provides a much richer data model than relational databases. Designed as a high-performance NoSQL database long before the term was conceived, it expresses data as a relational database, NoSQL database, object database, XML database, name-value pair database, document database, and more.

Characterized by its performance and scalability in concurrent access, the Ensemble database provides the ideal platform for bringing together and curating the heterogeneous set of data required for a data line-of-sight solution.

Analytics

Ensemble provides an analytics engine for real-time and on-demand insight into the information stored in the database and external sources.

Users can analyze cubes implemented directly onto operational data without the additional ETL (extract, transform, load) stage typically used to populate an external business intelligence tool. Incremental and continuous data loads into the cubes mean that analytics are based on the latest information, not data from last night or last month.

In addition to analyzing the structured data typically stored in databases, Ensemble can analyze unstructured or free-text data in the database and extract meaning for use in analytics. The same functionality can extract meaning from other sources, such as internal communications and market news, to "fingerprint" events relevant to your risk and compliance operations. A simple example is correlating market news to changes in stock trading volume.

By extracting meaning in this way, Ensemble can give personnel such as compliance officers insight into all of the organization's data, not just what is stored in traditional databases.

Data access: Adapters

Without a data line-of-sight solution, the biggest challenge is getting data from different silos into a common format to feed into existing analytics or compliance engines. Because Ensemble has already normalized the data formats and terminology of different silos into one coordinated feed, it can present analytics and compliance engines with a clean, single form of the data. The same adapters, transformation tools, business rules engine, and mapping tools used to bring data into Ensemble from many sources are used to feed data from its internal, canonical form out to existing applications in the correct format.

In the course of processing this data, Ensemble's internal analytics engine generates dashboards based on the raw data or maintained state, and executes business rules to recognize exceptions and initiate actions or alerts.

Data access: Business process management and human workflow management

Ensemble's workflow engine operates in conjunction with business process management to include people in automated business processes. The workflow engine can add requests to a user's work list based on username or role and wait for task completion, such as approving or verifying the request.

Security

Ensemble keeps all data and processes secure based on authentication, authorization, auditing, and database encryption.

- Authentication mechanisms supported include Kerberos, LDAP, Passwords, delegated authentication (for custom authentication mechanisms), and operating-system-based. Support is built in for two-factor authentication.
- Authorization determines what each user is allowed to use, view, or alter. For simplicity, users are assigned roles and each role determines what privileges belong to that user. To temporarily give users additional privileges (such as permission to modify a database), roles can be assigned to an application, and that application can then elevate the user's privileges while it is being used.

- For secure auditing, Ensemble records all system and application events in an append-only log, which is compatible with any query or reporting tool that uses SQL.
- Encryption of data at rest (on disk) includes indexes, not just the data itself, using the Advanced Encryption Standard (AES). Journal files can also be encrypted.
- Encryption of data in motion uses SSL and its successor, TLS to secure connections of various types.

Operations

Governance

For operational governance, Ensemble includes encryption of data at rest and at all points over the wire, user and role-based access control, strong authentication, and usage reporting and auditing. Ensemble also can plug into existing security and data governance environments like Guardium, BMC Patrol, and SNMP-based technologies.

Ensemble's ability to maintain extensive metadata is key to effective data governance. The system tracks what the data is, how it was created, how it has been modified, who has access to it, and how and when it has been accessed.

High availability

Ensemble database mirroring provides high availability and disaster recovery to minimize downtime after a system failure, and to allow business continuation after a data center failure.

Ensemble does not require downtime for routine maintenance and thus keeps scheduled downtime to a minimum. Use of database mirroring also greatly reduces the time required for software upgrades, including upgrading the operating system or the version of Ensemble.

Scalability

Ensemble can support high levels of concurrent access and large data volumes. Horizontal scaling with nothing shared is available for on-premises or cloud installations, and horizontal scaling where data is shared between nodes is available through Enterprise Cache Protocol (ECP), InterSystems' highly optimized caching protocol that allows the sharing of database files between nodes.

Conclusion

The Holy Grail of total operational awareness is much closer with a data line-of-sight platform. We have outlined a process and the technology components necessary for such a solution. Whether you decide to build or buy a data line-of-sight platform, pay particular attention to scalability, data management performance, business process management, and analytics, including linguistic analysis. InterSystems Ensemble is based on data management and integration technology proven in financial services. One customer's global equity platform, for example, handles 15% of the world's trading volume, has supported a billion trades per day, and requires half as much hardware as the previous implementation. As a complete and unified solution, Ensemble eliminates the need to purchase and integrate disparate components on your own. It leaves your organization with the time to focus on crucial risk and compliance activities, rather than on creating the IT platform to support them.

Take the next step

Call us to discuss your data line-of-sight goals and how we can help. InterSystems has worked with thousands of customers to help them deploy Ensemble within their organizations, on-premises, or in the cloud, or to embed it as part of their own software products.

- InterSystems.com/financial
- 1-800-753-2571 or +1-617-621-0600

See InterSystems.com/contact for a link to the number of your local office.

¹ "Five global banks to pay \$5.7 billion in fines over rate rigging," by Karen Freifeld, Reuters, May 20, 2015.

² "InterSystems Recognized as a Leader in Gartner Magic Quadrant for Operational Database Management Systems," Oct. 21, 2014.

InterSystems Corporation
World Headquarters
One Memorial Drive
Cambridge, MA 02142-1356
Tel: +1.617.621.0600
InterSystems.com

