talend

Data Integration Checklist

Technical Note

The need for data integration tools exists in every company, small to large. Whether it is extracting data that exists in spreadsheets, packaged applications, databases, sensor networks or social media feeds, there is a significant benefit to share and reuse information instead of having duplicate processes and silos of information. It is also important to select a solution that can address all your data integration needs, whether it be data integration, data migration, big data integration, data warehouse integration, or integration with business intelligence systems.

The following checklist provides key functional requirements for implementing and deploying data integration in an enterprise environment. Use the list to validate and prioritize your needs.

	Included	Description
Connect and Deliver		
Connect to Traditional Data Sources		Connect to data stored in relational databases, OLAP applications, non-relational structures like flat files, XML, common packaged applications like SAP, cloud-based applications such as salesforce.com, semi-structured (e.g Excel) data, unstructured (e.g. audio, video) data, and messaging systems. Support for industry standards like EDI.
Connect to Big Data and NoSQL		Integration with big data technologies (e.g. Hadoop, Hbase, Hive), big data platforms (e.g. Cloudera, Hortonworks, MapR) and NoSQL databases (e.g. MongoDB, Cassandra)
Data Movement		The ability for data consumers to receive data in many ways. Support bulk data movement, data services, data federation, change data capture (CDC) and direct data replication between data sources.
Data Synchronization		Support Extract Transform and Load (ETL) and Extract Load and Transform (ELT), real-time delivery, and event-driven delivery (trigger or changed data).

Transformation

Simple Transformations		Such as calculations, data type conversions, string manipulations, aggregations, automatic lookup and replace operations.
Advanced Transformations		Such as slowly changing dimensions, normalization of data, advanced parsing capabilities and transformation to complex standards (EDIFACT, HL7, and others)
Custom Transformations		Ability to create new custom transformations, as well as extend existing transformations.
Enrichment		Solution should have the capability to use enrichment data from a wide variety of sources. Enrichment data might come in various file formats and schemas both internal and external. It may come from online sources through service APIs, commercial partners or data providers.

Development and Data Modeling

Single Product		Support for all data delivery and integration operations, from connect, transform and load, via a single product
Graphical Tooling		Easy-to-use, graphical, drag-and-drop tools to build processes and transformations, and design data models, metadata and data flows. Graphical representation of objects and connectors. Wizards to automate common tasks.
Business Model Tooling		A non-technical tool that enables collaboration between technical and business users to structure all relevant documentation and technical elements supporting the data integration process.
Data Model Creation		Ability to create and maintain data models. Use graphical

and Management		tools to define relationships.
Metadata Management		Provides automated discovery of metadata. Ability to search metadata across multiple sources and show its lineage. Use a single repository of metadata across all product features, with the ability to seamlessly share and synchronize metadata between data integration tools and other tools (e.g. data quality, data profiling and master data management).
Business Rules and Workflow		The ability to define and manage business rules and execution flows. Process execution can be scheduled immediately, at a set time, or based on an event.
Versioning		Developers can easily version metadata, routines, processes, transformations or any other object used in the integration process. Then have the ability to see changes and roll-back to a prior version if necessary.
Collaboration		A set of tools for each user, i.e. business users, developers, and IT operations staff; and a shared repository consolidating all project information and enterprise metadata shared by all stakeholders.
Testing, Debugging and Tuning		Tools to test processes with data in the graphical tool, then interactively debug and tune for optimum performance.
Impact Analysis		Use graphical tools to **compare** processes, **assess the impact of change** and **view data lineage** to see where changes occurred.
Standards Support		To facilitate ramp-up time and leverage existing resources, products should embrace standards such as Eclipse, Java, JDBC, ODBC, and Web services.
Reusability		Should be able to reuse projects, metadata processes, cleansing, validation, enrichment and other highly used

		routines in a fast and easy manner.
Customizable		Generated artifacts can be customized for maximum flexibility. Ability to create your own custom components. Easy to customize and extend transformations.
Data Governance		
Integration with data quality tools		Integrated functionality with tools that profile and cleanse data, parse and standardize data, and match, merge and identify duplicate records to then be rationalized based on your requirements. The ability to define business rules to be applied to data.
Integration with data profiling tools		Integrated functionality with tools that do column-based analyses, dependency analyses, trend analyses and custom analyses.
Integration with MDM tools		Integrated functionality or out-of-the-box integration with tools to create a unified view of information and manage that master view over time.
Reports and Dashboards		Pre-built and customizable reports that show key data quality metrics over time. Provide the ability to export results in a variety of formats including XML, PDF, HTML, etc. Provide a dashboard (web-based) reporting system of data quality metrics and provide metadata to business intelligence (BI) systems.
Deploy		
Multi-Platform Runtime Support		Ability to seamlessly deploy to Unix-based, Linux-based and Windows systems. Ability to run on-premises and in the Cloud and virtualization environments. Ability to run in big data (MapReduce) distributed processing environments. Ideally generates code for portability and

		performance.
Load Balancing and Scalability		Clustering capabilities to spread server load over several machines. The ability to handle very large data volumes, working with big data and multi-terabyte data warehouses.
Failover		Ability to rollback a transaction and continue processing if there is a server failure without losing data.
Remote Execution		Ability to run processes remotely on various operating systems using the same configuration
Data Integration Services		Ability to deploy all aspects of run-time functionality as services within a service-oriented architecture.
Middleware Compatibility		Integrated functionality with MOM and ESB systems
Hadoop Support		Deploy native MapReduce jobs directly to a cluster with no needed appliances or additional software installed on the cluster. Have the ability to scale MapReduce processes with the cluster without code changes.
Monitor and Manage		
Centralized Administration		Ability to monitor and manage all resources and deployments from one location.
Web-based Monitoring		Ability to monitor resources and deployments from any browser.
Reports and Dashboards		Pre-built and customizable reports that show key data integration metrics. The dashboard shows information and statistics over time, e.g. performance, load volumes, subtask individual metrics such as database read and rights or enrichment service response times.

Exception Reporting and Management		Ability to define, report and handle exceptions when they occur. Capability to invoke special processes when violations to data integration rules. Examples include an e-mail alert, text message, or halting a process.
Security Controls		A mechanism to secure in-flight messages between applications as well as user/role-based security in the tool itself, LDAP.
Business User Interaction		Solution should provide an easy-to-use environment for business users to follow the key performance indicators for data integration, e.g. PDF reports and web-based portals.
Cloud Support		Ability to setup, deploy, and shutdown a cloud instance, e.g. Amazon EC2. Enables you to expand your computing capacity for your integration processes.
Support		
Comprehensive Support		Provides the support you need when you need it, e.g. community forums, web knowledge-base, email support, and phone support. 24x7 mission critical support. SLAs for response time, bug fixes and maintenance updates.
Training		Classroom, Online, and On-demand training for newbies, advanced developers and administrators.
Professional Services		Vendor offers a complete spectrum of consultative services: assessment, strategy and architecture, quickstart, design and development, tuning, technical audit, and custom offerings.

Talend Data Integration

Talend Data Integration provides an extensible, highly-scalable platform to access, transform and integrate data from any business system in real time or batch to meet both operational and analytical data integration needs.

With over 800 connectors, Talend connects natively to databases, packaged applications (ERP, CRM, etc.), SaaS and Cloud applications, mainframes, files, Web services, data warehouses, data marts, and OLAP applications.