ACTIAN™
a division of HCLSoftware

EBOOK

Data Engineering Guide: Nine Steps to Selecting the Right Data Integration Tool

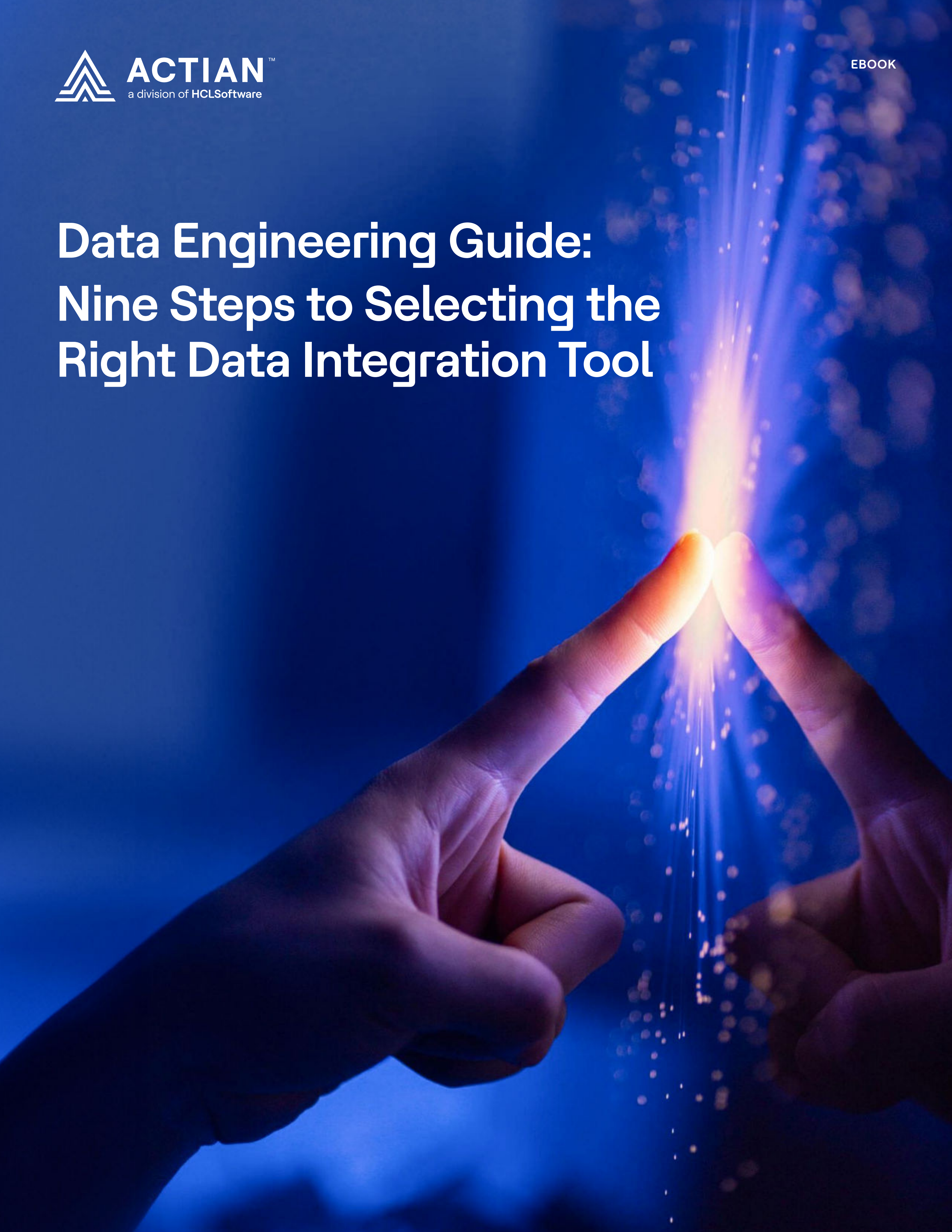


Table of Contents

Data pipelines are the foundation for high-impact data projects like business intelligence dashboards. But standing up a data pipeline can feel like an enormous project—and in many cases, it is. As data engineers plan their extract, load, and transform (ETL) technology stack, there are several key considerations and questions for them to keep in mind.

With organizations using an average of **130 apps**, the problem of data fragmentation has become increasingly prevalent. As data production remains high, data engineers need a robust data integration strategy. A crucial part of this strategy is selecting the right data integration tool to unify siloed data into a single source of truth in a data warehouse or data lake. This guide will walk you through the nine essential steps and key considerations in evaluating and selecting a modern ETL tool.

1. Assessing Your Data Integration Needs

Before selecting a data integration tool, it's crucial to understand your organization's specific needs and data-driven initiatives, whether they involve improving customer experiences, optimizing operations, or generating insights for strategic decisions.

Understand Business Objectives

Begin by gaining a deep understanding of the organization's business objectives and goals. This will provide context for the data integration requirements and help prioritize efforts accordingly. Collaborate with key stakeholders, including business analysts, data analysts, and decision-makers, to gather their input and requirements. Understand their specific data needs and use cases, including their specific data management rules, data retention policies, and data privacy requirements.

Audit Data Sources

Next, identify all the sources of data within your organization. These may include databases, data lakes, cloud storage, SaaS applications, REST APIs, and even external data providers. Evaluate each data source based on factors such as data volume, data structure (structured, semi-structured, unstructured), data frequency (real-time, batch), data quality, and access methods (API, file transfer, direct database connection). Understanding the diversity of your data sources is essential in choosing a tool that can connect to and extract data from all of them.

Define Data Volume and Velocity

Consider the volume and velocity of data that your organization deals with. Are you handling terabytes of data per day, or is it just gigabytes? Determine the acceptable data latency for various use cases. Is the data streaming in real-time, or is it batch-oriented? Knowing this will help you select a tool to handle your specific data throughput.

Identify Transformation Requirements

Determine the extent of data transformation logic and preparation required to make the data usable for analytics or reporting. Some data integration tools offer extensive transformation capabilities, while others are more limited. Knowing your transformation needs will help you choose a tool that can provide a comprehensive set of transformation functions to clean, enrich, and structure data as needed.

Consider Integration with Data Warehouse and BI Tools

Consider the data warehouse, data lake, and analytical tools and platforms (e.g., business intelligence (BI) tools, data visualization tools) that will consume the integrated data. Ensure that data pipelines are designed to support these tools seamlessly. Data engineers can establish a consistent and standardized way for analysts and line-of-business users to access and analyze data.

2. Choosing the Right Data Integration Approach

There are different approaches to data integration. Selecting the right one depends on your organization's needs and existing infrastructure.

Batch vs. Real-Time Data Integration

Consider whether your organization requires batch processing or real-time data integration—they are two distinct approaches to moving and processing data. Batch processing is suitable for scenarios like historical data analysis where immediate insights are not critical, and data updates can happen at regular intervals, while real-time integration is essential for applications and use cases like Internet of Things (IoT) that demand up-to-the-minute data insights.

On-premises vs. Cloud Integration

Determine whether your data integration needs are primarily on-premises or in the cloud. On-premises data integration involves managing data and infrastructure within an organization's own data centers or physical facilities, whereas cloud data integration relies on cloud service providers' infrastructure to store and process data. Some tools specialize in on-premises data integration, while others are built for the cloud or hybrid environments. Choose a tool that depends on factors such as data volume, scalability requirements, cost considerations, and data residency requirements.

Hybrid Integration

Many organizations have a hybrid infrastructure, with data both on-premises and in the cloud. Hybrid integration provides flexibility to scale resources as needed, using cloud resources for scalability while maintaining on-premises infrastructure for specific workloads. In such cases, consider a hybrid data integration and data quality tool like Actian's **DataConnect** that can seamlessly bridge both environments and ensure smooth data flow to support a variety of operational and analytical use cases.

3. Evaluating ETL Tool Features

As you evaluate ETL tools, consider the following features and capabilities:

Data Source and Destination Connectivity and Extensibility

Ensure that the tool can easily connect to your various data sources and destinations, including relational databases, SaaS applications, data warehouses, and data lakes. Native ETL connectors provide direct, seamless access to the latest version of data sources and destinations without requiring custom development. As data volumes grow, native connectors can often scale seamlessly, taking advantage of the underlying infrastructure's capabilities. This ensures that data pipelines remain performant even with increasing data loads. If you have an outlier data source, look for a vendor that provides Import API, webhooks, or custom source development.

Scalability and Performance

Check if the tool can scale with your organization's growing data needs. Performance is crucial, especially for large-scale data integration tasks. Inefficient data pipelines with high latency may result in underutilization of computational resources because systems may spend more time waiting for data than processing it. An ETL tool that supports parallel processing can handle large volumes of data efficiently. It can also scale easily to accommodate growing data needs. Data latency is a critical consideration for data engineers because it directly impacts timeliness, accuracy, and usability of data for analytics and decision-making.

Data Transformation Capabilities

Evaluate the tool's data transformation capabilities to handle unique business rules. It should provide the necessary functions for cleaning, enriching, and structuring raw data to make it suitable for analysis, reporting, and other downstream applications. The specific transformations required can include data deduplication, formatting, aggregation, normalization etc., depending on the nature of the data, the objectives of the data project, and the tools and technologies used in the data engineering pipeline.

Data Quality and Validation Capabilities

A robust monitoring and error handling system is essential for tracking data quality over time. The tool should include data quality checks and validation mechanisms to ensure that incoming data meets predefined quality standards. This is essential for maintaining data integrity and accuracy, and it directly impacts the accuracy, reliability, and effectiveness of analytic initiatives. High-quality data builds trust in analytical findings among stakeholders. When data is trustworthy, decision-makers are more likely to rely on the insights generated from analytics. Data quality is also an integral part of data governance practices.

Security and Regulatory Compliance

Ensure that the tool offers robust security features to protect your data during transit and at rest. Features such as SSH tunneling and VPNs provide encrypted communication channels, ensuring the confidentiality and integrity of data during transit. It should also help you comply with data privacy regulations, such as General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA).

Ease of Use and Deployment

Consider the tool's ease of use and deployment. A user-friendly no-code or low-code interface can boost productivity, save time, and reduce the learning curve for your team, especially citizen integrators who can come from anywhere within the organization. For example, this can be a marketing manager who wants to integrate web traffic, email marketing, ad platform, and CRM data into a data warehouse for attribution analysis.

Vendor Support

Assess the level of support, response times, and service-level agreements (SLAs) the vendor provides. Do they offer comprehensive documentation, training resources, and responsive customer support? Additionally, consider the size and activity of the tool's user community, which can be a valuable resource for troubleshooting and sharing best practices.

4. Considering Cost

Budget constraints are a significant factor in selecting a data integration tool. Consider the following cost aspects:

Subscription and Usage Costs

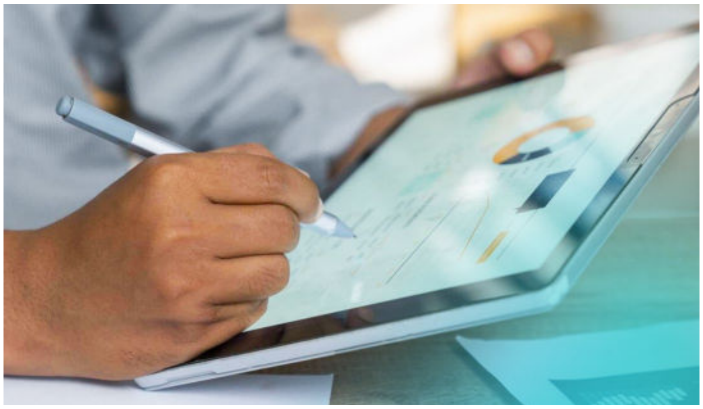
Evaluate the pricing model of the tool, whether it's based on a per-user license, data volume, or other factors. Calculate the total cost of ownership (TCO) over time to ensure it fits within your budget. While automated data pipelines offer significant benefits, they have associated recurring costs for every row of data moved from the source to the data warehouse. A solution like **Actian Data Platform** with unified data integration, quality, and warehousing in a single platform can help organizations that demand cost control and predictable cloud pricing.

Infrastructure Costs

Take into account the infrastructure costs associated with the tool. This includes development environments, hardware, cloud services, network, and maintenance expenses. Organizations incur both capital (CapEx) and operational (OpEx) expenses in a hybrid integration model, depending on the components and resources used.

Ongoing Maintenance and Support Costs

Consider the long-term costs of maintaining and supporting the data integration tool. Cloud providers handle infrastructure maintenance, including updates, backups, and high availability, reducing the operational burden on organizations. In a hybrid model, on-premises components require in-house maintenance and upgrades.



5. Testing and Proof-of-Concept

Before making a final decision, it's essential to test the shortlisted data integration tools to ensure they meet your requirements.

Creating a Test Environment

Set up a dedicated test environment that mimics your production environment. Use sample data to simulate real-world scenarios. A well-designed test environment allows data engineers and analysts to validate data transformations, troubleshoot issues, and evaluate pipeline performance without impacting live data.

Data Integration Testing

Using the selected tools, execute various data integration tasks, such as data extraction, transformation, and loading. Prepare various scenarios, including common and edge use cases. This data should closely resemble production data and have sufficient volume to test scalability.

Performance Testing

Conduct performance testing to assess how the tools perform under varying workloads. Conduct load testing by gradually increasing the volume of data to assess how the data integration processes handle increased workloads. Measure performance metrics such as response times and resource consumption as data volume scales.

6. Gathering Peer Feedback

Involve database administrators, application developers, data scientists, BI developers, and data analysts in the evaluation process to gather feedback on the usability and effectiveness of the data integration tool.

Involving End Users

Seek input from members of the data team who will be using the tool daily. Data engineers should work closely with cross-functional teams to implement the integration strategy effectively and mitigate risks associated with data movement.

Addressing User Concerns and Feedback

Address any concerns or feedback provided by end-users during the evaluation process. This ensures that the tool aligns with their needs, expectations, and use cases. Their insights can help identify any user experience issues or specific requirements such as data residency and compliance requirements.

7. Selecting a Vendor

Based on your assessments, it's time to select a vendor and finalize your choice of data integration tool.

Shortlisting ETL Tools

Narrow down your choices to a shortlist of data integration tools that best align with your organization's requirements and budget constraints. Industry reports, recommendations from peers, and online research can help in this process.

Requesting Trials

Request free trials or demonstrations from the shortlisted vendors. This allows you to see the tools in action and verify their capabilities. If a free trial is unavailable, conduct a proof-of-concept (POC) or pilot project with shortlisted vendors.

Evaluating Vendor Reputation and References

Research the reputation of the vendor and seek references from their existing customers. Look for case studies and customer reviews. A data vendor with a strong track record and positive customer testimonials is more likely to provide reliable support and updates.

8. Implementation and Data Migration

Once you've selected a data integration tool, it's time to plan and execute the implementation.

Planning the Implementation

Create a detailed implementation plan that includes timelines, resource allocation, and a step-by-step approach to deploying the chosen tool. Develop rollback and contingency plans and ensure that there is a way to revert to the previous state or recover from errors without significant disruption.

Data Migration Strategy

Develop a data migration strategy to move existing data from various sources to the data warehouse while ensuring data accuracy, completeness, security, and minimal disruption to business operations. Determine what data needs to be migrated and what data can be left behind or archived.

Data Quality Assurance

Implement data validation and quality assurance processes to verify that the integrated data in the warehouse meets the desired standards. Set up data quality rules to promptly address any data discrepancies or pipeline issues.

9. Monitoring and Optimization

Continuous monitoring and optimization are essential to maintain the efficiency and reliability of your data integration processes.

Setting Up Monitoring

Implement monitoring tools and dashboards to track the performance of data integration jobs. Set up alerts for potential issues. These rules are essential in data engineering pipelines to identify and rectify data issues, anomalies, and errors before it is used for analysis, reporting, or other downstream processes.

Performance Optimization

Regularly review and optimize data integration workflows for efficiency. Adjust configurations and resources as needed to ensure optimal performance. Utilize parallel processing, batch sizing, data partitioning, and indexing techniques to monitor and adapt to changing data and infrastructure needs.

Continuous Improvement

Encourage a culture of continuous improvement within your data team. Gather feedback from users and stakeholders—particularly those managing the data warehouse and BI tools to identify areas for enhancement.

Outcomes You Can Expect with Actian

Selecting the right data integration tool is a critical decision for data engineers tasked with providing the foundational infrastructure required for analytics and machine learning (ML) initiatives. You can make an informed choice by following the steps outlined in this guide and carefully considering your organization's unique needs.

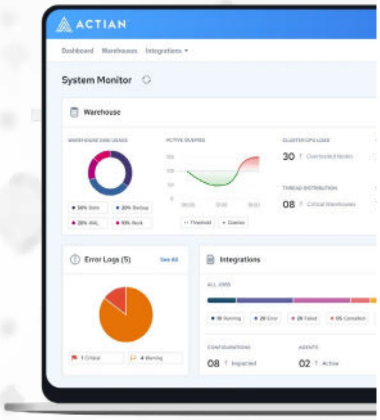
A fully managed hybrid data platform like the one from Actian simplifies complex data integration challenges and gives you the flexibility to adapt to evolving data integration needs. With Actian, you can get started with **DataConnect**, a standalone, hybrid data integration and quality tool with more than 300 connectors or **Actian Data Platform**, a unified data integration and warehousing platform. Either way, you'll be confident knowing you have the toolkit to work with a variety of data sources and formats to maintain reliable data pipelines with ease.

The best way for data engineers to get started is with a free trial (insert tracking URL) of the Actian Data Platform. From there, you can load your own data and explore what's possible within the platform. Alternatively, book a demo (insert tracking URL) to see how Actian can help automate data pipelines in a robust, scalable, price-performant way.

Get Started Now

DataConnect ↗

Actian Data Platform ↗



About Actian

Actian makes data easy. We deliver cloud, hybrid cloud, and on-premises data solutions that simplify how people connect, manage, and analyze data. We transform business by enabling customers to make confident, data-driven decisions that accelerate their organization's growth. Our data platform integrates seamlessly, performs reliably, and delivers at industry-leading speeds. Learn more about Actian, a division of HCLSoftware: **www.actian.com.**

Tel +1 512 231 6000 Fax +1.512.231.6010 710 Hesters Crossing Road, Suite 250, Round Rock, TX 78681 **Actian.com**

ACTIAN™
a division of **HCLSoftware**