



WHITEPAPER

Transitioning from Manual to Automated Data Pipelines

Table of Contents

Introduction 3

What Is a Data Pipeline? 4

How Manual Data Pipelines Work 4

Moving Beyond Manual Data Pipelines 5

Automated Data Pipelines as a Transformative Solution 7

Success Stories: How Businesses Have Leveraged Automation in Data Processing 9

Key Elements Empowering Data Pipeline Automation 12

Future Trends in Data Pipeline Automation 14

Conclusion 16

Introduction

[According to Statista](#), global creation and consumption of data reached 64.2 zettabytes in 2020 and is expected to grow to more than 180 zettabytes by 2025. While data continues to grow, structures and sources, it becomes more and more difficult to integrate it into a unified view, process it and transform it into insights. Yet, companies demand swift and precise insights based on this complex data.

The evolving intricacies and the sheer magnitude of data are driving businesses away from manual data handling and towards automated data pipelines. This transition from legacy processing to automated data pipelines becomes imperative to meet escalating data demands effectively. By automating data processing pipelines, businesses unlock speed, precision, scalability and cost efficiency in operational and analytical processes alike.

What Is a Data Pipeline?

Data pipelines function like arteries, carrying data from its various origins to destinations where it can be used for analysis, reporting or strategic decision making. Varying with the use case, it may involve different steps, such as copying data from on-premises storage to the cloud, integrating data from different systems, standardizing it etc.

A data pipeline is an intricate framework orchestrating every step of the data journey such as extraction from multiple sources, transformation into usable formats, and ultimately loading into storage or analytical systems.

To accomplish this data journey, data pipelines entail a suite of components. The components in a data pipeline can differ based on the use case. They may include resources for extracting data from its sources, sometimes authentication methods for accessing the sources, data processing components and resources to write data to the desired destination. They may also include defined workflows governing the sequence of tasks within the data pipeline project and robust monitoring mechanisms to ensure reliability.

A data pipeline's lifecycle spans from conception to operation, involving various stakeholders—enterprise architects, data engineers, analysts, and data scientists—who collaborate at different stages. It encompasses activities like gathering business requirements, architecting pipeline designs for quality and consistency, implementation, operational management, consumption of data outputs, and ongoing improvements guided by feedback and governance protocols.

How Manual Data Pipelines Work

Data and analytics are the cornerstone of effective business decisions. However, for many companies, this is still a theoretical aspiration because existing data consumption and transformation processes remain ad-hoc and manual.

In a typical use case of data pipeline where a business needs to analyze its sales data, transaction data will need to be transformed and moved to a data warehouse. From the data warehouse, another processing pipeline may be required to create aggregates and visualize the data in a BI tool, such as Power BI or Tableau.

Executing such projects using manual data pipelines relies heavily on human intervention in each step. Companies require a team of data engineers to write code specific to each use case. Tasks such as data extraction, transformation, and loading are performed manually by individuals or teams using custom-coded scripts or tools. Data engineers oversee the data movement manually, which is time-consuming and error prone. These pipelines are typically built on an ad-hoc basis to fulfill specific, immediate needs, often lacking standardized processes, flexibility, scalability, and automation.

Moving Beyond Manual Data Pipelines

Manual data pipelines come with their inherent limitations. Some of the most common challenges associated with manual data pipelines include:

Data Volume and Variety

The volume, variety and speed at which data is being produced is constantly growing. With the multitude of data generating technologies, there's a rapid influx of data available to businesses. For instance, monitoring systems continuously generate log files, IoT devices generate huge volumes of streaming data, and the numerous financial transactions being recorded all contribute to this influx. Data is available in diverse formats and can originate from different kinds of sources, including structured sources like relational databases and Excel files, semi-structured sources like delimited files and JSON files, and unstructured sources such as PDFs.

This exponential rise in different aspects of data, collectively known as big data, presents both opportunities and challenges to data stakeholders. While there is plenty of raw data available, reporting and decision-making requires refined, analytics-ready data.

Data pipelines facilitate the conversion of raw data into usable, analytics-ready data. However, the escalating volumes and complexities of data present significant hurdles for conventional, ad-hoc data pipelines in achieving this goal. Handling large data volumes is resource-intensive, requiring time, effort and expertise. Increasing demands can quickly exhaust resources, resulting in performance bottlenecks, inefficient use of skilled personnel and increased operational costs.

Data Quality Issues

Human involvement in data transformations leads to frequent errors, including typos, formatting inconsistencies, and flawed logic within scripts or code. The use of varying methodologies among pipeline designers can lead to non-uniform data processing, ultimately resulting in analyzing inconsistent data and misleading conclusions.

Integrating data from disparate sources manually can lead to complexities in data mapping. Inconsistencies in the representation of data across sources require careful attention and time from technical resources to implement integrations without damaging the data quality.

Lack of Scalability

Ad-hoc manual pipelines lack flexibility and scalability. They're not designed to accommodate the evolving demands of growing data volumes and diverse data sources. Scaling manual processes becomes impractical and often requires extensive rework or redevelopment.

Lack of Governance

Manual data pipelines are typically built without standardized procedures or oversight. This lack of governance results in difficulties in maintaining data integrity, tracking changes, ensuring compliance, and establishing clear ownership of data assets.

These challenges highlight the inability of manual data pipelines to cope with the expanding data landscape. Where manual data pipelines fall short in addressing these escalating complexities and demands, there's a need to find more advanced, adaptive solutions.

Automated Data Pipelines as a Transformative Solution

The landscape of data handling is witnessing a transformative shift with the emergence of automated data pipelines. This evolution is a response to the escalating complexities in data volumes, structures, and the demand for rapid insights. Automation in data pipelines shifts the pipeline development process away from manual, ad-hoc approach towards sophisticated systems that efficiently manage all aspects of data workflows. These encompass design, development, testing, deployment, orchestration, scheduling, and monitoring.

In automated data pipelines, the repetitive tasks involved in each of these implementation stages are performed by software instead of humans. Some of the key advantages of this transition from manual data pipelines to automated data pipelines include:

Reduced Time-to-Insights

Automation expedites data processing, enabling quicker extraction, transformation, and loading (ETL) processes. This swift data-to-results timeline empowers faster decision-making and enhances operational agility.

For example, [Vonage](#), a communication provider, which was initially observing an average provisioning time of 4 days was able to bring down the time to 4 minutes. By automating their data pipelines, Vonage was able to process data from multiple sources, like customer transactions and product catalogs, create quotes and automatically send them to customers.

This drastic reduction in processing time demonstrates how automation streamlines data handling, enhancing the company's capability to respond promptly and efficiently to evolving business needs.

Scalability

Automated pipelines offer elasticity, allowing for dynamic adjustments to fluctuating data volumes and demands. This scalability ensures seamless handling of increased data flow without the need for extensive manual intervention or system reconfiguration.

Precision and Accuracy

Automation reduces the potential for human errors inherent in manual handling. Consistently applied predefined rules and processes in the automated dataflows enhance data accuracy, ensuring reliable outcomes and consistent analytics.

Cost-Effectiveness

Automated pipelines reduce reliance on specialized expertise for routine tasks. For manual data processing, businesses typically rely on trained employees who can implement data loading tasks to move data from the source to the destination, in addition to writing complex SQL scripts to transform and integrate source data. Using automated solutions to implement data pipelines eliminates the need to write SQL scripts or design the data loading process by hand. This, in turn, minimizes the need for skilled resources, optimizing operational costs and resource allocation.

Reduction in Human Intervention

By automating scheduling, triggering, and monitoring, the need for constant human oversight decreases. By reducing the time required to be spent on managing data, employees can spend their time efficiently on more creative tasks. With accurate insights as an outcome of automated data pipelines and the spare time to utilize these insights, they can focus more on understanding customer requirements, strategizing solutions and taking up innovative initiatives.

The emergence of automated data pipelines is transforming data management approaches, addressing the limitations of manual processes while offering enhanced efficiency, accuracy, and adaptability in handling and processing data.

Success Stories: How Businesses Have Leveraged Automation in Data Processing

Numerous examples of business achieving transformative goals through data processing automation outline almost every industry.

Some success stories that stand out are:

NETFLIX

Netflix, a global entertainment giant, has undergone a significant transformation in its data management approach by embracing automated data pipelines.

Use Case

Netflix's vast data ecosystem involves handling diverse data sources for content recommendations, user engagement analysis, and infrastructure optimization.

Traditional Approach

Previously, managing their data workflows involved manually constructing deployment pipelines. However, as the complexity of their data operations grew, so did the challenges. Issues arose with versioning, transparency, and maintaining ACID deployment standards. Rollbacks to previous workflow versions became intricate tasks requiring meticulous asset reconstruction and realignment.

Transitioning to Automated Solution

To address these challenges, Netflix developed Dataflow, a CLI tool for comprehensive data pipeline management. This tool automated the asset deployment process, providing a robust abstraction layer over the manual deployment design. Leveraging Dataflow, Netflix engineers could seamlessly deploy, manage, and version assets and workflows, eliminating complexities related to runtime dependencies and fixed storage locations of code assets. [Automated data pipelines](#) also established a standardized approach to data handling as opposed to ad-hoc implementations, simplifying onboarding of new team members, improving consistency across pipelines, and minimizing human errors as codebase evolves.



[Ever.Ag](#) is a leading ERP software provider in the agricultural and food supply chain domain.

Use Case

Netflix's vast data ecosystem involves handling diverse data sources for content recommendations, user engagement analysis, and infrastructure optimization.

Traditional Approach

Traditionally, the company would extract data from its sources manually and write it to the destination, consuming valuable time and resources.

Transitioning to Automated Solution

Seeking a scalable remedy, they decided to switch to automated data pipelines. By designing, deploying and orchestrating data processing pipelines in Astera's Data Pipeline Builder, Ever.Ag was able to streamline data importing tasks, bringing down processing times to mere minutes. Without relying on IT teams, they can implement field mappings seamlessly, even for varying customer-specific data formats. Translating customer-specific codes, which conventionally required writing complex lookup queries, are now easily implemented by business users using pre-built lookup blocks in a no-code environment.

Automating data pipelines not only accelerated data processing for Ever.Ag, but it also optimized their overall organizational efficiency.



Another example of adopting data pipeline automation to streamline their workflows is that of Michigan Education Association (MEA). It is a self-governing education association, supporting over 12000 teachers and other education-related professionals throughout the state.

Use Case

To continue supporting education professionals, MEA needs to utilize data to resolve member issues, process invoices and organize bargaining units.

Traditional Approach

Before transitioning to automation in data processing, MEA required a team to download and unzip folders containing data in different formats, such as CSV, PDFs and flat files. Data from individual files needed to be extracted manually, processed using ad-hoc pipelines and loaded into a SQL database. Handling files containing millions of records was challenging and time-consuming. It also necessitated the physical presence of the employee at work to download the files and execute the integration process.

Transitioning to Automated Solution

Adopting Astera Data Pipeline Builder for automating their data processing lifecycle allowed them to eliminate the repetitive manual work and speed up the process. After designing the data flows, workflows and execution schedules in a no-code environment, the software could handle the rest of the work without requiring any human intervention.

Without requiring the presence or time of an employee, the entire process from downloading data from multiple files to processing and loading to the database at user-specified intervals or triggers is handled by the machine alone. Scheduling the integration process to be executed automatically overnight, MEA employees were able to access analytics-ready data in their target database at the start of each workday with no effort from their side.

Key Elements Empowering Data Pipeline Automation

Automated data pipelines involve several key components, each contributing to the successful implementation of self-sustaining data operations.

These include:



No-Code/Low Code Pipeline Design

No code and low code solutions are an important part of data pipeline automation today. These solutions have not only empowered business users to design and execute data pipelines, but also simplified it for technical users. Users with varying technical backgrounds can use these solutions to construct robust pipelines without writing extensive codes from scratch. Utilizing intuitive drag-and-drop interfaces, these solutions offer an array of pre-built components that accelerate the pipeline development process.

By addressing the complexities associated with traditional coding, no-code/low-code tools facilitate rapid designing and deployment, significantly reducing the development lifecycle. With a standardized and modular approach to pipeline design, these platforms encourage collaboration, enabling business analysts, data engineers, and domain experts to contribute directly without heavy reliance on specialized programming skills.



Data Quality and Governance Framework

Ensuring data integrity, consistency, and compliance remains a critical aspect of automated data pipelines. Automated data cleansing, data validation and data quality checks ensure that the health of data is up to mark for the downstream applications it will be used in. By implementing these data quality checks at the required stages of the data pipelines, organizations can ensure that any errors arising in the data are resolved automatically and timely so the results can be accurate and complete.

Furthermore, modern data pipelines also make it easier to implement data governance. Companies can automate data governance processes, establishing clear ownership and accountability over data assets. Moreover, it fosters a culture of trust, enabling stakeholders to make informed decisions based on reliable, high-quality information.



Workflows

Efficient workflows and dependency management are fundamental in orchestrating data pipeline execution. Workflows define the logical order of tasks in the automated execution of data pipelines.

They ensure that tasks are executed in the correct order. Users can direct tasks to be executed sequentially, in parallel, or even alternatively based on the successful or unsuccessful execution of its prerequisites. This ensures streamlined execution of the end-to-end process, reinforced with user-defined error-handling, without necessitating human intervention at any stage.



Automated Execution Via Scheduling

Besides workflows, scheduling is another important component of automated execution of data pipelines. Scheduling mechanisms as part of automated data pipeline orchestration allow users to schedule pipeline execution either continuously, at pre-defined intervals, or in response to specific triggers.

Without requiring a human to trigger the pipeline manually, scheduling allows a variety of possibilities, including timely data updates, event-driven actions, and real-time processing actions. By scheduling, pipelines can be configured to run during off-peak hours, reducing system load and optimizing performance. This ensures that data processing tasks are executed in a timely manner, facilitating faster decision-making and operational efficiency.



Monitoring and Alerting

Once the data pipelines are up and running, robust monitoring and alerting mechanisms are needed to ensure their health and reliability. Real-time monitoring allows for continuous tracking of pipeline performance, identifying bottlenecks, and proactively addressing issues that may arise during execution.

Monitoring, error-handling and alerting functionalities can also be automated to minimize human intervention in this area, while ensuring timely corrective measures. Alerting functionalities provide notifications for anomalies, errors, or deviations from predefined thresholds, enabling timely intervention.

In response to such notifications, pipeline engineers can perform strategic optimization techniques. These techniques may involve fine-tuning pipeline configurations such as enabling pushdown optimizations to convert the transformation logic to SQL queries to be implemented on the target system. It may also involve optimizing server resources by reallocating or load balancing across multiple servers for greater efficiency.

These elements collectively contribute to the effectiveness, reliability, and scalability of automated data pipelines, empowering organizations to derive valuable insights from their data with minimum resources and skills.

Future Trends in Data Pipeline Automation

The generation and utilization of data is continuously growing, propelling data pipelines into a future brimming with innovation. The endless possibilities in data pipeline automation and the impact it can create for businesses is beyond exciting. Here's a peek into some promising trends that can be seen emerging in the future:

AI-Driven Orchestration and Maintenance

Artificial intelligence has already set foot in almost every application, and it's no different for data pipelines. In the foreseeable future, AI will become the backbone of data pipelines, orchestrating operations and maintaining them with unparalleled precision and adaptability. As industries adopt AI-driven data pipelines, they will be able to intelligently manage complex processes like task allocation, resource utilization, and workflow optimization. Machine learning algorithms dynamically allocate tasks and resources, adjusting execution paths in real-time based on workload priorities and historical performance data. These systems can predict potential bottlenecks or failures, rerouting tasks or scaling resources to ensure optimal efficiency without manual intervention.

Beyond orchestration, AI capabilities can also be extended to deployed pipelines to introduce proactive monitoring, diagnostics and self-healing capabilities in the pipeline. These systems continuously monitor pipeline behavior, detecting anomalies or deviations in performance. Utilizing predictive analytics, AI predicts potential issues before they occur, enabling proactive maintenance or corrective actions.

Real-Time Data Pipelines

The imminent rise of real-time data pipelines promises lightning-fast analytics fueled by continuous data streams. Businesses will leverage instant insights derived from a combination of the freshest and historical information, enabling swift, strategic decision-making. Real-time data pipelines work by continuously ingesting data streams from multiple data sources. The data is continuously transformed, integrated and loaded for analytics. Such pipelines can enable businesses to make instant decisions and take strategic actions based on real-time metrics.

To be able to leverage real-time data processing, businesses need to ensure certain capabilities in their data pipeline solution. Stream-processing capabilities, parallel processing, Change Data Capture, API consumption features, and high-speed ingestion from multiple sources concurrently are some of the features to look for in the automated data pipeline solution to leverage real-time insights.

Adoption of DataOps

The data renaissance will also witness an increased adoption of DataOps, an approach combining agile development, DevOps and lean manufacturing principles. It is igniting a transformative wave in data pipeline automation. Picture an era where collaboration, efficiency, and innovation converge to revolutionize how data flows and fuels business decisions.

DataOps unifies diverse technical and non-technical teams to contribute to the data management lifecycle. Eliminating silos, enhancing cross-functional cooperation, and adapting automation, DataOps ultimately aims to maximize the value of data for the organization.

Strategic implementation of DataOps, employing a set of best practices, diverse skillsets from within the organization and futuristic technologies can empower businesses to embrace a data driven culture. Data-first companies like Netflix, LinkedIn, and Uber, have already adopted DataOps, spear-headed by continuous integration / continuous deployment (CI/CD) principles to skyrocket data-quality and optimize deployment. Google is already evolving the concept of DataOps into MLOps, which is an application of CI/CD principles to machine learning models.

In the coming years, DataOps will be approachable for more companies by employing self-service data pipeline automation tools. These tools will facilitate collaboration among users of varying technical skillsets, streamline workflows, and ensure consistency. From CI/CD pipelines for data pipelines to observability and monitoring components, the right tech stack is pivotal to the transformation of the data culture.

Conclusion

The transformation from manual to automated data pipelines is driven by the need for speed, precision, and scalability, all of which are indispensable for businesses to adapt to the complexities and volumes of modern data. By addressing limitations associated with manual processes, automated data pipelines empower organizations to make data-driven decisions more efficiently.

These pipelines are easily scalable, enhance precision, are cost-effective, reduce human intervention, and – most importantly – expedite data processing, drastically reducing time-to-insights. This acceleration empowers faster decision-making and enhances operational agility in an increasingly competitive landscape.

The transition to automated data pipelines isn't merely an upgrade; it's a strategic imperative. The faster businesses can seek out pipeline automation technologies, the faster they can position themselves for success in an era defined by insights, innovation, and adaptability.



Contact us for more information or to request a free trial
sales@astera.com | 888-77-ASTERA
www.astera.com

Copyright © 2023 Astera Software Incorporated. All rights reserved.
Astera and Centerprise are registered trademarks of Astera Software
Incorporated in the United States and / or other countries. Other
marks are the property of their respective owners.