tray.io

WHAT YOU NEED TO KNOW

Reverse ETL goes mainstream

TABLE OF CONTENTS

Introduction

The Extract, Transform and Load (ETL) process and the tools that enable it were popularized in the 1990s with the growth in data warehousing. Way back, enterprises started using ETL tools like Informatica to map and extract and merge data from operational systems and loading into a data warehouse for reporting and analytics.

ETL has been around for a long time, and it's been through more than a few iterations over the period, with evolutionary steps like Extract, Load, and Transform (ELT). And the Sources and Targets have certainly changed dramatically, too—CRM, ERP, HCM, and other systems across cloud and on-premises, and evolving databases for the warehouses themselves.
But the approach has fundamentally remained the same—suck data out of Marketing, Sales, Finance, HR, Service, and other apps, and load it into a database designed for data analysis, with transformations along the way, prepping that data into an appropriate schema, so it's optimized for fast queries.

However, the last few years have seen perhaps one of the most significant changes since its inception, Reverse ETL. It's where data isn't just flowing into data warehouses like Redshift, Snowflake for analytics; it's also flowing out of them to drive (and improve) business processes themselves. So the data that was meant for driving better decision-making can also drive better automated processes and more standardization of data across the stack.

It's how marketing organizations, who have been leading the charge in reverse ETL, are reinventing business processes like personalization, customer engagement, lead management, fueled with data from the data warehouse. They're making these processes more intelligent and more responsive, often using Redshift, Snowflake, BigQuery, Azure SQL, and other traditionally data warehouse-oriented database platforms to get there. In some ways, it's a testament to the flexibility of modern cloud databases. For example, databases like Snowflake have morphed from supporting customer analytics to enabling Customer Database Platform (CDP) use cases—where they have now essentially become a single source of the truth for crucial customer data.

"...marketing organizations, who have been leading the charge in reverse ETL, are reinventing business processes like personalization, customer engagement, lead management, fueled with data from the data warehouse."

Modern data warehouses like Redshift and Snowflake have rich APIs and support high-performance query processing that can elastically scale. So it's only natural that their data begins to be used in high-frequency business processes like lead routing or order management processes, rather than just analytics.

Where this evolution has profound implications is traditional ETL tools themselves. They were designed for the ETL use case—operational Sources to analytical Target. Reverse ETL flips the equation. Operational apps become the Target, and the warehouse becomes the Source, changing integration patterns, workloads, needs, and even the users of these tools. For example, Reverse ETL may require injecting data on-demand into business processes as they happen or inform conditional logic in those workflows. Or it may be used to sync data from the warehouse out to a broad range of apps, changing the standard data flow from many sources to one target, to instead one source (the warehouse) driving many targets.

A brief history of ETL

What's important to remember about traditional ETL is that these tools are essentially designed for a one-way data flow, from one or many sources to one or a handful of targets. Operational systems, like SAP ERP, or PeopleSoft HCM, might be the Source systems. And a database like Oracle, Azure SQL Data Warehouse, Redshift, BigQuery, or Snowflake, where the data warehouse resides, is the Target. There may be other Targets for ETL, such as a Data Lake, like Azure Data Lake, Qubole, or AWS Lake Formation. In this case, data is left rawer, less structured, and processed less than its data warehouse counterpart—less useful for business users, but a richer data set that data scientists will love.

And there's one more, in the last few years, the target might also be a Data Lakehouse, pioneered by Databricks, which reduces the need to manage both a Data Warehouse for business users and a Data Lake for data scientists. A Lakehouse can support both BI and SQL workloads, together with data science and data engineering patterns in a single platform. But whether it's a Data Warehouse, Data Lake, or Data Lakehouse, ETL tools are designed to flow data from Source systems into them, the Targets, ultimately for analytics and data science.

An evolution of ETL is also Extract, Load, and Transform (ELT). The transformation process (deduping, cleansing, enriching, aggregating, etc.) is often highly processing-intensive when moving data. With modern database platforms, it's often more efficient to perform the transformations once the data has landed. So, ELT tools will first extract and load the data into the data warehouse and then complete the transformations once the data has landed there, rather than doing it along the way.

A whole ecosystem of vendors grew up around ETL, and more recently, ELT. Such as Informatica, Talend, and more recently, tools like Matillion, and Fivetran, but ultimately, ETL/ELT tools are usually designed to facilitate that flow of data from operational Source systems, like CRM, ERP, HCM, SCM, mainframe, and whatever else, into a warehouse, data lake, or lakehouse.

Many older ETL tools are notoriously complex and expensive to maintain, designed for hard-core data integration specialists, making it painful for business users to exercise control. However, some more modern ETL tools provide a degree of support for citizen integrators to build their data warehouse.

It's important to remember that ETL tools are most often used on a scheduled, batch data integration basis and usually aren't designed for triggered operational business processes. So, for example, you'll rarely want to use an ETL tool to perform an Order-to-Cash process, which might be triggered when an Order is closed in Salesforce, and that single record then loaded into an ERP like Microsoft Dynamics GP. While higher-end ETL tools may support Event Stream (real-time) processing, it's, however, they're still often engineered around flowing data from operational Source systems to Target platforms for analytics.

"Many older ETL tools are notoriously complex and expensive to maintain, designed for hard-core data integration specialists, making it painful for business users to exercise control."

The Emergence of Reverse ETL

One of the first areas where Reverse ETL is seeing significant growth is in Marketing. Because it turns out that traditional data warehouses like Redshift, Snowflake, and BigQuery can work incredibly well as a CDP. They can serve both customer analytics requirements while also acting as the customer master to help drive marketing processes and help standardize data across the dozens of apps in a modern marketing stack.

In Marketing, using Snowflake or Redshift as your CDP, the data can be used for driving Lead Routing, Email Personalization, Lead Scoring, and countless other processes, in addition to enabling prospect and customer analytics. It can also be used as the single source of the truth to keep all the other apps in the marketing stack up to date with the latest customer data. Rather than maintaining point-to-point integrations, data can be synced out of Snowflake or Redshift—and into Marketo, Eloqua, Salesforce, HubSpot, and any other component of the marketing stack, enriching and standardizing their data.

For example, FICO was an early adopter of this approach in Marketing. Rather than building point-to-point data integrations to keep all of their apps in sync (which ultimately leads to a web of integrations and complex data syncs that must be maintained), they switched to Reverse ETL. Instead, they're using their existing Redshift deployment as their CDP by using the Tray Platform to flow data from their data warehouse across their stack. That way, they have a single up-to-date source of the truth, flowing master data into their apps.
FICO synced Marketo from their data warehouse to pull lead and customer data directly from it, allowing them to consolidate and dedupe their marketing lead lists automatically. They also allowed the team to create a single source of truth to inform marketing preferences within Marketo and manage their customer data and engagement data from Redshift. The FICO team ETLs this data to a table in Redshift and then Reverse ETL's it to provide regular intraday data updates to Marketo.

Using Reverse ETL has enabled some significant business process improvements. According to FICO's Consumer Marketing Manager, "Our data is cleaner, fresher, and we've been able to leverage that in our communications, from nurture campaigns to triggered campaigns. Implementing Reverse ETL and automation has improved deliverability for our campaigns from approximately 60% of emails landing in inboxes to about 90%. Our engagement rates spiked overall with open rates increasing by about 26% and click-through rates by about 9%."

"Implementing Reverse ETL and automation has improved deliverability for our campaigns from approximately 60% of emails landing in inboxes to about 90%. Our engagement rates spiked overall with open rates increasing by about 26% and click-through rates by about 9%."

**— Consumer Marketing Manager
FICO**

Thinking Beyond ETL: Reverse ETL Requires Different Integration and Automation Capabilities

This blurring of what a data warehouse is—not only as a database for analytics but also to drive better business processes, requires rethinking what to expect from an ETL tool. It's no longer enough for it to simply load the data warehouse because that's now only just half the equation.

You're probably thinking, do I need one tool for ETL and another for Reverse ETL? The answer is nuanced. It depends. Perhaps you already have a traditional ETL tool in place and a production data warehouse. In that case, the chances are you'll almost certainly need another solution for Reverse ETL if you want to retain your existing ETL tool for loading the warehouse. Start by examining the use cases. For example, can it be used as your source for master data to standardize your prospect, customer, product, suppliers, or other data across your stack? Or will it be leveraged to make actual business processes more intelligent, like adding more personalization to email marketing campaigns on-demand? Is there an opportunity to use it to upgrade scoring and routing based on intent data or as your enriched account master to drive better ABM marketing?

Most ETL tools fall short because they have limited event-based triggers, are designed for integration, not automation, and require specialists to build integrations.
Using highly trained integration specialists to build ETL processes works when you have a handful of integrations, e.g., managing the handful of flows to load your data warehouse(s). But when you plan to use the warehouse to fuel dozens of business processes across your business, that approach simply won't scale—and where citizen integration and low-code tools come into play.

Alternatively, if you want to start with a clean sheet of paper, some modern integration and automation platforms enable ETL and Reverse ETL in the same toolset. These platforms can simplify skill sets, aid reuse, and make maintenance easier too. If you're familiar with ETL capabilities, you'll need to ensure the platform supports some more capabilities so it's suitable for Reverse ETL too.

Critical capabilities to enable Reverse ETL

Ensure your Reverse ETL tool supports event-based triggers and webhooks

Most ETL tools don't support Triggers and Webhooks, but they're not designed to automate business events as they happen, but these are crucial capabilities for Reverse ETL. Traditional ETL tools are primarily designed for bulk data integration, not event-based automation.

For example, with a Reverse ETL use case, when an event happens in Salesforce or Marketo, you might want to trigger a workflow that might query your data warehouse for some data to inform lead flow automation. In many modern cloud applications, Webhooks can enable this, where the application calls out to a Webhook URL when a particular business event occurs. Webhooks can provide a lot of flexibility as so many business apps support them, so you'll want your integration/automation platform to respond to them too.

Additionally, ensure your integration platform natively supports built-in application triggers, like when an opportunity changes to an order in Salesforce or a ticket is created in Zendesk. Integration and automation platforms that connect to those events can enable even more ways to trigger a workflow that can, in turn, accesses data from your warehouse and can perform an action based on it.

Data integration to handle both use cases

If you're looking for a single platform to cover data in both directions, load your warehouse, and sync your apps with it, you'll want to ensure that it can handle core data integration, both to load your DW and extract from it on-demand.

Start by ensuring it can map any field easily using drag and drop and support multi-step data flows that combine, aggregate, conditionally process, transform and enrich any data. Ensure there are helpers to quickly transform, enrich, and convert any data type, field, from text, date/time to numeric. Built-in data storage can be a huge help to perform lookups, denormalize and aggregate data, create lists and dimension tables, all within a workflow, or share data across workflows.

Built-in CSV transformation can aid in importing (or exporting) CSVs, joining, sorting, updating, parsing them. While business logic to handle loops, conditional logic, and scripts can help with deduping, concatenating, and refactoring data in both directions.
While this kind of functionality is often built into ETL, it's usually reasonably technical to get

at—it's essential to ensure citizen builders who are creating Reverse ETL processes can get value out of it too.

Plan for a material difference in ETL and Reverse ETL workloads

With ETL loading a data warehouse, it's all about bulk large-scale data volumes as you move millions of rows (or much more) of data in. If you're evaluating a single platform for ETL and Reverse ETL, you'll want to be sure it scales for both use cases.

For loading the warehouse, ensure it can process and enrich large volumes using batching, pagination, parallelization, and other strategies.

But Reverse ETL workloads are different; they're event-based and high frequency, often with much smaller amounts of data per integration execution (think integration with a marketing process querying your data warehouse thousands of times per second). Or they may require syncing with a large volume of apps across the stack simultaneously (like syncing master data from your DW to many marketing apps).

Conventional ETL tools aren't designed for these kinds of workloads, while traditional process automation tools often aren't built for bulk data movement.

One strategy is to use an integration platform based on a more modern serverless architecture, which is elastic at the core, versus older conventional platforms that require sizing, provisioning, care, and feeding, which is hard to do when you have mixed workloads. With a serverless approach, compute to drive bulk or event-driven workloads is dynamically scaled on-demand. In tandem with parallel processing, it's an approach that can provide the backbone to support both workloads in one solution.

> **"With a serverless approach, compute to drive bulk or event-driven workloads is dynamically scaled on-demand. In tandem with parallel processing, it's an approach that can provide the backbone to support both workloads in one solution."**

You'll likely need more and different connectors for Reverse ETL

Most ETL tools are light on connectivity, as they have just enough connectors to fuel common data warehouse needs and load them. While they certainly support all the mainstream data warehouses, like Oracle, Snowflake, Redshift, BigQuery, and provide connectors to many of the sources often needed to drive data needs for analytics, like common CRM, ERP, HR sources, etc., they often don't have the broad connectivity to sources that aren't typically used to load warehouses.

For example, many systems aren't typically the sources for ETL (e.g., lead routing, lead scoring, data enrichment, collaboration tools, event management, etc.), so ETL tools don't have connectors to them. But in a Reverse ETL case, these are potentially significant targets that would often benefit from pulling data from the warehouse or used in tandem with it in business process automations.

So, whether you're going with one platform for ETL and Reverse ETL, or complementing your existing ETL with another integration platform, ensure that it has all the connectors built-in for current and future use cases. It's also helpful that your platform for Reverse ETL includes an HTTP connector for any REST, SOAP, or GraphQL endpoint, to avoid being painted into a corner when a connector doesn't exist or doesn't support the needed endpoint out of the gate.

Recognize that your Reverse ETL builders will build differently

There's one more thing. It's essential to recognize that the folks who will be building Reverse ETL integrations are often close to the business processes themselves in their departments. It's roles like marketing ops manager, growth marketers, or RevOps leaders, not your classical ETL specialists. These users might be looking to lead the charge in adding data warehouse data to improve a lead routing flow or better email personalization. That's different from an ETL data integration specialist's specific expertise or focus.

> **"These users might be looking to lead the charge in adding data warehouse data to improve a lead routing flow or better email personalization."**

These roles will likely be productive with a different, more collaborative low-code experience than traditional data integration platform tooling can provide. A more diagrammed business process-oriented experience can make development easier and faster and reduce the need to rely on hard-to-find developers (and free them up to work on other projects). Instead, business users are looking to easily compose automations, assembling integrations using business logic that consists of conditionals, loops, branching, nesting, and error-handling, visually tapping into events, apps, database, or other workflows.

For this requirement, it's better to try before you buy. Not every app that's pitched as low-code is easy. At the same time, other low-code tools compromise flexibility (which means you'll eventually need developers to get the project over the finish line.) So get business teams hands-on, see how fast and easy it is for them to build automation while tapping into their data warehouse and whether it'll work with your Reverse ETL use case(s). `
To start trying out integrating apps and business processes to your data warehouse, check out a Tray.io trial, or get a personalized demo.

Reverse ETL goes mainstream

+1.415.418.3570 | EMAIL | WEBSITE | BLOG

About this guide

Written by

Paul Turner, Tray.io

REVERSE ETL GOES MAINSTREAM
VERSION 1.0 | 13.08.21