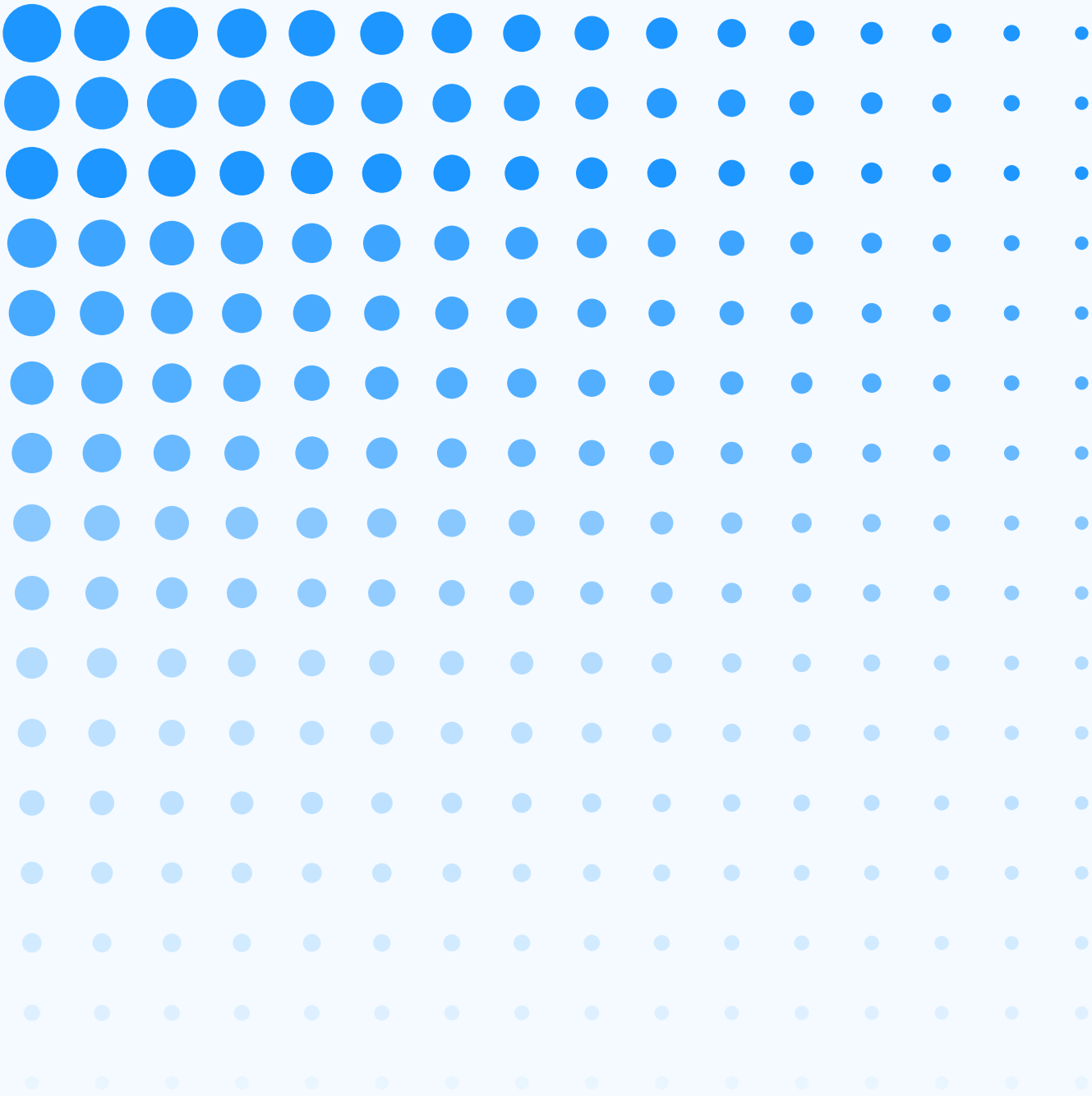
In-House or Cloud-Native ETL?

HOW TO MAKE THE RIGHT CHOICE



integrate.io

Table Of Contents

01

INTRODUCTION

From the moment that businesses began using computers, they've faced the complex problem of moving data from A to B.

It's hard enough when you have a direct one-to-one connection that syncs data between two systems. This requires you to think about things like scheduling, data integrity, compatibility and security.

But the problem gets even trickier when you have a one to many connection, such as when multiple systems are feeding data to a central repository. The logistical issues become exponentially more difficult every time a new system comes online.

In this era of analytics-driven business, this age-old problem is now a high-stakes challenge. Often, a business will turn to a dedicated kind of data management platform known as Extract, Transform, Load, or ETL.

ETL performs three discrete tasks in sequence:

Extract: First, the ETL queries a data source, such as an enterprise system. It uses that source's unique interface to obtain the data it needs.

Transform: After that, the ETL puts the data through some transformations to make it compatible with the destination database. These transformations often happen on a dedicated database known as the staging layer.

Load: Finally, the ETL pushes the transformed data towards its ultimate destination. In doing so, it may create tables or merge with existing data.

When an ETL is set up and running automatically in the background, it creates a data pipeline, allowing data to flow smoothly from A to B.

ETL has been around in various forms since

the early days of enterprise software. Many early systems were hand-coded by in-house developers using the available tools, and some companies still take this DIY approach today. Others buy an out-of-the-box in-house ETL solution that they can install on their network.

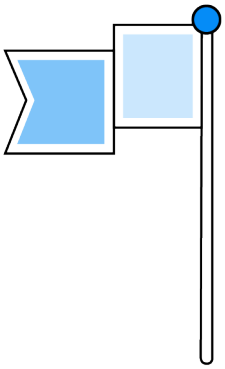
If you're trying to optimize your data infrastructure, you may find yourself choosing between an in-house and cloud-native solution. You might also wonder if it's time to upgrade your existing data pipeline.

In recent years, cloud-native ETL has emerged as the optimal method for connecting disparate data sources. Cloud ETL is fast, secure, cost-efficient and ready for a world that runs on the Cloud.

In this document, we're going to look at the differences between an in-house ETL and the cloud alternative. We'll study the pros and cons of each, and you'll find out how to pick the best solution for your data infrastructure.

02

DEFINING YOUR CRITERIA FOR A SUCCESSFUL ETL PROJECT



Before you dive into the available ETL options, it's essential to consider why you need it in the first place.

ETL implementation can be a major change project, so it's important to work with stakeholders to outline factors like:

- The current state of your data infrastructure
- Success metrics
- Project constraints

Let's explore what all of this means in terms of an ETL project.

The Current State of Your Data Infrastructure

You'll generally start by studying your enterprise's current state to get a better idea of where things need improving. This stage might involve steps like:

- **Mapping existing data infrastructure**: Get a bird's eye view of every data system in your stack. Take notes of data-generating systems (sources) and data storage systems (destinations). Look at how these systems handle data and interact with each other. If you don't already have an ETL, you may find that you have a mishmash of manual and automated synchronization processes.
- **Identifying pain points**: When you review your current infrastructure, you'll likely notice that there are some inefficiencies that cost time or money, or you may discover problems that threaten data reliability. By understanding these pain points, you'll be able to identify areas where an ETL process can help deliver value.
- **Reviewing your data governance policy**: This policy should describe your organizational approach to maintaining secure and accurate databases. The rules in this policy will outline what you can and can't do with an ETL project. If you don't

- already have a clear data governance framework in place, it's a good idea to establish one as soon as possible. Speak with your CTO, CIO, or DPO for clarity to ensure whichever ETL tool you choose, takes measures to prevent data breaches.

Once you know where you currently stand, you can start thinking about a future with ETL.

ETL Success Metrics

What does a successful ETL project look like? Your business might focus on any of the following outcomes:

- **Data quality**: Any data infrastructure project should prioritize data quality above all else. Many companies invest in ETL because their current solution creates issues such as data loss, duplication or inconsistency. Your ETL plan should help meet key quality metrics, which you can monitor with ongoing data audits.
- **Data availability**: People and processes need timely access to data. For instance, if a customer buys something through e-commerce, your ETL needs to make that data immediately available to the order fulfillment system. Customer service will also need the data so they can field any queries. You can set a performance target here by establishing what kind of update frequency you need for each database.
- **Analytics quality**: ETL can help gather all of your data in a central repository, which makes it easier to perform detailed analytics. You can measure these outcomes by working with your analytics team to see if they have the data they need, when they need it. If your ETL project is a success, it should enable your analytics team to provide meaningful insights.
- **IT support overheads:** Your local IT team will always have to provide some support, but your goal is to reduce this as much as possible. You can ask them to track calls related to data integration. After deploying your ETL, those calls should decrease.

Once you know what you want to achieve, you'll be able to identify the right solution for you.

ETL Project Constraints

Every project exists within certain limitations that will influence your decision. Some of the key obstacles to consider here are:

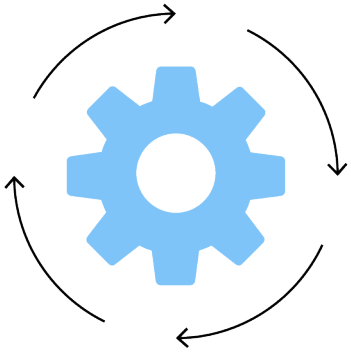
- **Cost**: IT projects often have hidden costs. As well as the sticker price of your chosen ETL software, you have to think about the cost of hardware resources, cybersecurity, support staff, training and maintenance.
- **Skills**: There's also the question of whether you have the in-house skills to deliver your chosen ETL solution. For instance, you might feel tempted by a DIY solution, but does your IT team know how to build, and more importantly, maintain one?
- **Flexibility**: In an architectural sense, ETL sits between the other systems in your stack. When you add or change a system, it may directly impact how your ETL performs. If you think you might change or grow your IT infrastructure in future years, you'll need to make sure you have a suitably flexible ETL solution.
- **Scalability**: There's also the issue of growth, which can lead to an exponential increase in the data volumes that pass through your ETL. Will your proposed infrastructure be able to deal with a sudden growth spurt?

At the end of this review, you should have a clear understanding of three things:

1. Where you are now
2. Where you want to get to
3. What obstacles might hold you back from your target

03

HOW AN IN-HOUSE ETL WORKS



In-house ETL is an automated Extract, Transform, Load system that runs on your hardware, within your network. This kind of service is not dependent on anything outside of your organization.

This approach has obvious advantages in terms of control, visibility and security. However, it raises other issues, such as how you interact with external systems and accidental data breaches from internal users.

There are actually two different types of in-house ETL. You can hand-code a DIY solution from scratch, or you can install an enterprise ETL system from a vendor.

Hand-coded ETL

IT teams sometimes build a data pipeline from scratch. They usually take this approach when the requirements are minimal, like when you need to link one or two data sources to a repository. Hand-coded ETL is often a temporary stop-gap solution, but it can become part of the infrastructure over time. The downside to this approach is the amount of time it takes to create and maintain the data pipelines.

Often, these solutions are not ETL systems in the literal sense of the word. An ETL is a dedicated platform that handles data transformation and integration. Hand-coded ETL is often a collection of scripts and small tools that you manage with a scheduler like Cron. Usually, you would set up ETL processes to run during the night when the network is quiet.

Modifying and maintaining this type of system is a challenge, especially if the original developers are unavailable. When you need to modify a connection or add a new integration, you'll need to dive right into the code level, or you may need to create a new job from scratch. Without a central ETL platform, your data repositories will end up with a one-to-many relationship with the rest of the network.

Enterprise ETL

Once their network grows beyond a certain point, most organizations need to deploy a dedicated ETL solution. This brings a lot more stability to the network and reduces the chances of data loss.

An in-house enterprise solution such as Oracle or Informatica allows you to create a system of one-to-one connections like so:

- Each source connects directly to the ETL
- The ETL manages all data transformations
- The data repository has only one connection, which is direct to the ETL

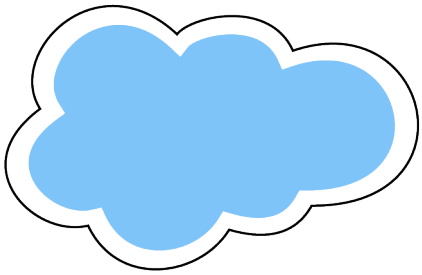
This lets you build a more modular infrastructure, in which each pipeline connection is independent of the other. You can manage all of your integrations from a central hub and view a single error report to keep track of any issues.

This kind of solution works like any other in-house enterprise software in that you need to provide sufficient hardware resources to host it. Once it's running, you have complete control over how it works. Unfortunately, you also have full responsibility for any issues that arise.

In-house ETL is great when all of your other systems are also in-house. If you need to connect to a cloud-native service such as a data warehouse, then you'll have to figure out the security and compliance implications of external data transfers.

04

HOW A CLOUD-BASED ETL WORKS



Modern companies tend to use a lot of cloud-based services. In a world of remote work and outsourcing, it makes sense to use the security and flexibility of the Cloud to connect with everyone.

As a result, you may have crucial nodes in your data network that sit outside of your network. Such systems include:

- A data warehouse or data lake
- CMS and marketing automation
- Enterprise Resource Planning (ERP)
- Finance and payroll systems
- eCommerce and digital tools

If you use an in-house ETL with an external endpoint, it raises some tricky security questions.

What if *both* endpoints are outside the network? Consider this common example: a business uses Shopify to provide an e-commerce portal. This data is integrated with information from other systems and stored on the company's in-house data warehouse. The company also uses the Cloud-native analytics service Looker for BI purposes.

With an in-house ETL, the data would have to go on a journey like this.

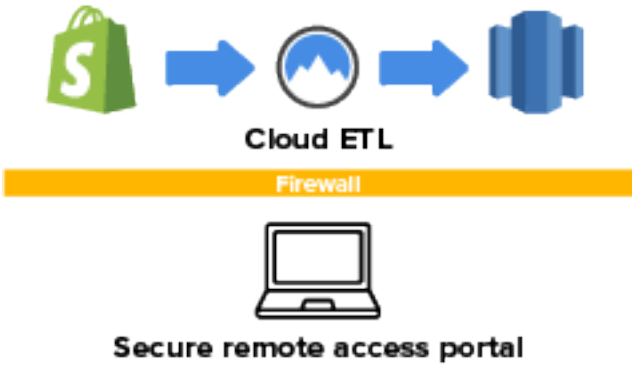
That involves crossing the internal firewall twice, which is slow at best and a security risk at worst.

Fortunately, you can now access powerful cloud-based ETL platforms. Like other as-

a-service solutions, a cloud-based ETL is hosted by a third party on their hardware. Setting up a cloud ETL service is as easy as opening a Netflix account. There's no need to provide a dedicated ETL environment or install any software. Everything runs on your service provider's infrastructure.

Once you initiate your cloud-ETL service, you access everything through a secure web portal. From here, you can create new integrations, manage data transformations and monitor data quality.

Now, your transaction looks like this:



This approach substantially reduces the risk to data within your network. It also helps protect external data, as cloud services have access to best-in-class security measures.

Features of an Enterprise-Grade Cloud-Native ETL

Cloud-native ETL platforms tend to have a better range of features than the in-house alternatives.

Take Xplenty as an example. Xplenty is the leading enterprise ETL for Cloud, and it offers some immensely useful features, including:

Easy integration: Integration is the most challenging aspect of data pipeline development. You and your team have to figure out how to connect to each individual source, and you then have to maintain that connection over time.

A platform like Xplenty includes a library of integration templates that you can use to connect to popular cloud services with only a few clicks. You can use the Xplenty generic REST API to connect to any other RESTful service.

A no-code/low-code interface: ETL integration typically requires at least two people. You need an IT expert who can write and execute the integration, but you also need a business expert who understands the data to decide where the data needs to go.

Modern ETL platforms make things easier by providing a simple drag-and-drop interface. Anyone can use these tools to create complex integrations between sources—even if they have limited development ability.

Best-in-class security: Cloud ETL providers go to enormous lengths to ensure data safety. The service is typically hosted in a secure physical environment with a SOC-2 compliant environment. Xplenty's service runs on Amazon AWS, the world's leading cloud provider.

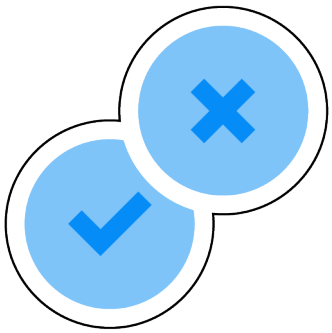
Xplenty offers additional layers of security, such as field-level encryption. This kind of encryption protects data from the moment it's extracted at the source until it arrives securely at the destination.

On-demand scalability: Most cloud services offer a range of membership tiers to suit individual needs. If you're growing, you can move up a tier. If you're experiencing challenging times, you can scale down. All of this happens in the background without impacting your service.

These are just some of the features that may impact your decision when considering your options. Let's dig deeper into the pros and cons of both.

05

IN-HOUSE VS. CLOUD: PROS AND CONS OF BOTH



Both of these approaches offer their own advantages and disadvantages.

Pros of In-House ETL

- **Control**: You have complete control over every aspect of your data pipelines, so you can configure the software as you please. If you're using a DIY solution, you can recode it from scratch.
- **Visibility**: You can track every individual bit as it moves through your network from source to destination. There are no blind spots in the process.
- **Ownership**: Your ETL process is your property. Nobody can cancel your service or revoke your license. If the vendor goes out of business, you can still use the software (although you won't have ongoing support).

Cons of In-House ETL

- **Hardware costs:** You have to provide the resources for your ETL, which might require buying and installing dedicated servers.
- **Lack of support**: If problems arise, it's your responsibility to deal with them. You'll need a skilled ETL support team that can resolve errors and maintain connections. You'll also need to deploy any new patches or updates.
- **Compatibility**: When you want to add a new service to your stack, you may find that it's not compatible with your existing ETL. This could mean that you have to write a new integration from scratch.
- **Security and compliance**: You are fully responsible for all security and compliance requirements of your ETL. If anything goes wrong, it's on your shoulders to resolve and recover. This not only means that you'll need in-house support, but you'll also need to keep

testing and watching for emerging threats.

- **Problems with scaling**: If your data activities become more intense due to demand, your ETL may struggle to keep up. You'll have to provide any additionally needed hardware, and your team may need to configure new integrations for the latest systems.
- **Never-ending upgrade cycle**: Every few years, you may find that you need to purchase the latest version of your ETL software. You will also need to keep updating the hardware that hosts your system.

Pros of Cloud-Based ETL

- **Ease of deployment**: Cloud services don't require local installation because everything is physically hosted on the Cloud. This cuts down on deployment time, plus you don't have the initial expense of providing hardware.
- **Ease of use**: A good cloud-based ETL platform like Xplenty offers a library of ready-to-use integrations. This allows you to connect to data sources and get an automated pipeline flowing, all with little or no coding.
- **Cloud compatibility**: Cloud ETL generally works well with other cloud systems, including data repositories like warehouses and lakes. They can also securely link back to your in-house systems, which gives you a great deal of flexibility.
- **Scalability**: A cloud system is massively scalable. When you need to process more data, simply switch to the next tier of service. Your cloud ETL provider can make more resources available to you, so you can scale up with zero disruption.
- **Security**: Cloud-based ETL providers typically offer cutting-edge security measures to protect their customers. Xplenty, for instance, gives you field-level encryption that safeguards data at every point of the journey.
- **Built-in service**: Your cloud ETL provider will provide a certain level of service. For example, they will handle all updates and security patches on their side, without disrupting your service. Some providers may offer additional support for more advanced ETL integration queries.

Cons of Cloud-Based ETL

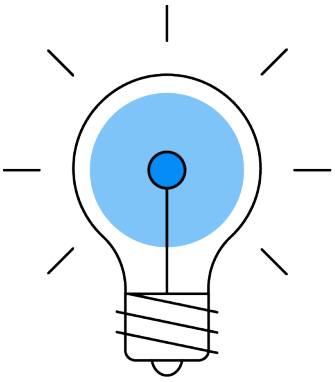
Lack of ownership: Cloud services run on a subscription model, which means you don't own any part of the system. If you breach the terms of your agreement, you could lose your account and all of its settings.

Ongoing fees: The subscription model also requires that you keep paying a fee for as long as you use the service. Over several years, this could potentially be more expensive than an in-house solution.

Lower visibility: Some ETL providers offer outstanding visibility into all data transactions. However, it will never be as transparent as your own solution.

06

HOW TO DECIDE BETWEEN IN-HOUSE AND CLOUD-NATIVE ETL



Advocates of in-house and cloud-native ETL platforms will each argue that their approach is the best.

In truth, neither one is the “best”—they’re both suited to different use cases. That’s why it’s important to set project goals and understand your project's constraints. This will help you find a solution that works for you.

To help you make the decision, here are checklists for both approaches:

In-House ETL Checklist

If you answer Yes to most of these questions, an in-house approach might suit you best.

- Do you need complete control over every aspect of data management?
- Are most of your enterprise systems located in-house (i.e., they are within your network)?
- Do you have the hardware resources to run your in-house ETL?
- Do you have the right people to deploy and configure your ETL?
- Can your IT team support ETL problems as they arise?
- Are you unlikely to need to scale up or scale down in the next few years?
- Do you have data governance rules that forbid external data transfers?
- Will your solution guarantee a high standard of data quality?
- Will your solution make data available where it’s needed, when it’s needed?
- Does an upfront cost suit your budget needs better than regular payments?

Cloud-Native ETL Checklist

If you answer Yes to the majority of these questions, you might consider a cloud-ETL approach.

- Do you need to minimize the amount of time your IT team spends on ETL configuration and management?
- Do you want to empower non-IT people to build and maintain integrations?
- Do you need an ETL you can deploy quickly and with minimal fuss?
- Would you benefit from the flexibility that allows you to scale with ease?
- Will you benefit from not having to provide in-house IT support?
- Is some of your data infrastructure already based in the Cloud? This can include production systems like Salesforce or ERP, data repositories like Amazon Redshift, or both.
- Do you have a team that often works remotely?
- Does your analytics team need faster access to quality data?
- Have you found a cloud ETL provider that provides a service within your budget?
- Can that provider guarantee data security and compliance at all stages of the ETL process?
- Will they provide ongoing support as your business needs evolve?

Making the Choice

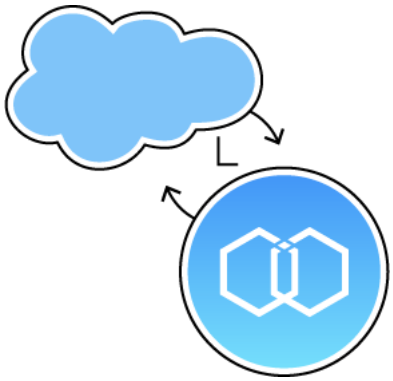
There are a lot of things to consider, including cost, security, compliance, scalability, and ease of use.

The one big factor that may influence your choice if you're starting from scratch is start-up costs. An in-house ETL has a substantial upfront cost--an expense you can't recoup if you make the wrong choice. Cloud ETL doesn't have that initiation cost, and many services offer extensive demonstrations and a free trial before you make a commitment.

But what if you're not starting from scratch, but instead considering moving on from your current solution? You may find yourself leaning towards a cloud ETL solution, but that would involve an upgrade project. In the next section, we'll talk about how to make the switch.

07

MIGRATING FROM IN-HOUSE TO THE CLOUD



Cloud migration can be painless if you prepare well and follow your plan. Your exact approach will depend on your individual data architecture. Still, in general, you will need to take these steps:

1. Perform a Full Impact Assessment

As with any change management process, you must map out the potential impact of any changes. This includes looking at all of the systems that are data sources or destinations. You'll also need to think about systems that depend on data from your ETL, such as analytics or business intelligence tools.

Remember to give equal consideration to the people who use the data. Talk to all stakeholders and ask about their pain points with the current process, as well as their must-haves for any new projects. If they're dependent on existing data pipelines, you have to try to minimize any disruption to their work.

2. Establish Key Performance Metrics

If you started out with a clear project definition, you should have a clear idea of your desired outcomes. These may include faster turnarounds for ETL processes, improved analytics quality or fewer calls to IT support.

Measure where you are right now and use this as a benchmark. You should also consider how you're going to measure any changes so you can objectively establish the level of value that your ETL creates.

3. Start With a Proof of Concept

It's often best to start with a small project, like integrating a single source into a standalone database within your data warehouse. If your first data pipeline is a success, you'll prove to the team that this migration will add value.

Alternatively, your pilot may not go so well. The advantage of a limited pilot project

is that you can test your strategy without causing any major disruption. A failed pilot offers valuable data, which you can use to improve, refine and get things right next time.

4. Focus on Cloud Systems First

Cloud systems are low-hanging fruit in any cloud migration. These systems are generally easy to integrate—you should be able to get a data pipeline flowing with very little code.

If you use Xplenty, you can start with services that are supported in the integration library. These ready-to-go integrations will allow you to connect to popular services like Salesforce, Google Analytics and Amazon S3 with just a few clicks. Once you've secured these easy wins, you can work with your IT team to integrate the more complex in-house systems.

5. Communicate Every Step of the Way

As you move ahead with your migration strategy, it's vital to keep everyone in the loop. Let them know how the project is progressing, and notify them in advance of any downtime. Remember to celebrate your wins and let everyone know when you've hit a major project milestone.

Communication is a two-way street, so you must provide a channel where people can provide feedback. User feedback is invaluable in a change project, as it will help you anticipate any issues before they arise. For example, your analytics team will be able to tell you if they see any data quality issues arising from the new ETL process.

6. Work With Your Vendor

Your ETL vendor knows their product better than anyone else, and they have experienced every possible use case. They're an essential resource to lean on, especially if you're trying to create a complex data infrastructure.

That's why it's a good idea to talk to your vendor beforehand and participate in some product demos. A live demo will give you a taste of what it's like to work with an ETL partner. You'll soon see if they speak your language and can cater to your needs.

7. Measure and Refine

Cloud ETL is often a set-it-and-forget-it proposition that doesn't require much ongoing attention. However, in the early stages of migration, it's important to monitor outcomes and see that you're on track to meet expectations. Pay close attention to data quality—you may want to perform regular spot checks and data audits.

If you're failing to meet your anticipated goals, go back and try to figure out why. Is it a configuration error? Is it a people issue that requires additional training and support? Now is the time to improve and refine your ETL so that it delivers what you need.

8. Dismantle Your In-House Solution

Dismantling an in-house system can take some work. You may have dedicated machines that you can repurpose or dispose of, and you'll need to cancel any software licenses that automatically renew. If the ETL process is a major part of someone's job, you'll have to reassign them.

Make sure that you're following a well-thought change management process when you take old systems offline. This way, you'll ensure continuity of service for everyone who is dependent on data.

08

XPLENTY: THE FUTURE OF ETL

Data is becoming more complex, which means that data integration is also becoming more complex. As your business moves through the 2020s, you'll find yourself growing more dependent on analytics. Meanwhile, your analytics team will need faster, better and more diverse data sources.

Xplenty is a cloud-native ETL that grows with you. It's a powerful solution that connects your enterprise systems to data warehouses, allowing you to create sophisticated transformations with just a few clicks.

Additionally, Xplenty supports more complex operations like ELT, ETLT and working with big data structures. It's your platform for success, no matter how your business grows.

To learn more, visit Xplenty today and organize a free demo with an Xplenty representative.