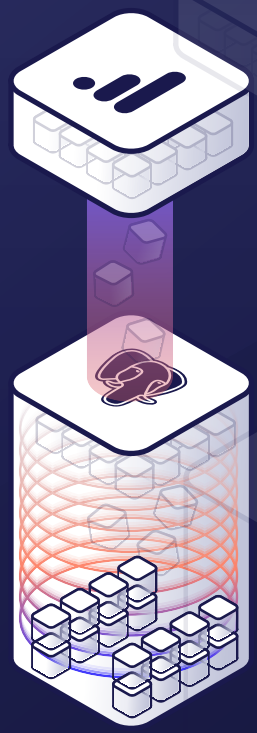

Building vs. Buying your ELT solution

Airbyte

Introduction

In the current economic context, your team is most likely asked to think about making the organization, including your data infrastructure, more efficient. Are you struggling to find the right ELT solution for your organization's growing data integration needs? Look no further than this comprehensive ebook. We will explore the options available to you, including building an in-house solution, using Airbyte Open Source, or utilizing Airbyte Cloud.

Through our detailed analysis, we will examine the features required to set up and maintain an effective ELT solution, as well as the associated costs. Whether you're a small business or a large corporation, we'll provide scenarios tailored to fit your specific organizational needs.

By the end of this ebook, you'll have a better understanding of the pros and cons of each solution, enabling you to make an informed decision and choose the best ELT option for your company. Don't settle for a subpar data integration strategy - let us help you find the right one for your organization.

How to navigate the ebook

To effectively navigate this ebook, it's important to understand the structure and organization of the content.

Firstly, the ebook is broken down into three main parts: setup, operation, and improvement. Each of these parts covers a different stage in the lifecycle of an ELT solution.

Secondly, for each part, we explore the options available to you, including building an in-house solution DIY, using Airbyte Open Source, or utilizing Airbyte Cloud. We discuss the actions that need to be taken for each option, including the estimated amount of work required, and the specific needs that each action item fulfills.

Finally, at the end of the ebook, we provide a summary of the costs associated with each solution, taking into account the size and needs of your organization.
We also built a tool to let you personalize the costs of each solution, by inputting the average engineering salary in your team, the number of connectors you need to build and maintain, and the overall volume of data you're looking to replicate every month.

Assumptions

Because Airbyte's focus is to build the best reliable data integration solution, we have a good understanding of what it takes to build the different components. The estimated engineering work is based on our experience.

Our estimations are based on the North American market, whether it is the level of skill of engineers, the salary, and organization's needs. Please don't hesitate our estimation tool to add your own estimations.

Variables:

- Airbyte's State of data survey found that on average, Data Engineer total comp is $211,000. To bring this to the actual cost for the company, we added a 30% for taxes, insurances, benefits and perks. Bringing the total cost of a Data Engineer to $274,300/year.
- Number of connectors will be 20, with 2 that are not yet supported by Airbyte, in that ebook. Some surveys showed that companies relied on 5 to 20 internal and external data sources. That said, the average Enterprise relies on 400 connectors.
- Infrastructure cost to run the data pipelines will be about 20% of Airbyte Cloud's price, as this is what we see.

Most companies already have some ELT components built in-house. That's why we built our estimation tool to be modulable, so you can select/unselect whatever work you're already done.

From 0 to Production

In this part, we will discuss what it takes to go from a blank page to a production-ready ELT solution so that data can start flowing between your sources to their destination.

DYI

Solve ELTs engineering challenges

By going the in-house route, you will first need to gather a good understanding of how an ELT actually works and the engineering challenges that need to be addressed to extract, load and transform data at scale. You have to ensure that the data is moved in an efficient, fast, reliable, and secure manner. While a proof-of-concept – where data moves from source to destination – can be quickly done, that is only the tip of the iceberg. There are so many other technical challenges to be solved; here is a non-exhaustive list:

- How to handle schema changes?
- Does it monitor and alert for data sync failure?
- Can it handle future additional data connectors?
- How do you handle incremental synchronization?
- Are you supporting batch and real-time ingestion?
- How to build a data pipeline that can be distributed?
- What happens if a data source is temporarily down?
- How does the system recover when it crashes in the middle of a data transfer?
- What happens if the amount of data going through the system increases by 10×?

This means that even before building the pipeline, you will most likely want to add the following features:

- Orchestration and automated scheduling
- Incremental updates and possibly Change Data Capture support for databases.
- Automated schema migration
- Idempotence (recovery from failure) and lossless replication

Build data connectors

Once the core of your ELT solution is in place, you will need to build data connectors that need to handle both the source and destination of the data to be moved. This can be a complex and time-consuming process, as each connector will need to be customized to fit the technical specifications of the data source or target, as well as handle any data movement and transformation requirements that are dictated by the business needs. The transformation step can include filtering, aggregating, or enriching data as it moves through the pipeline. This will ensure that the data validation, clean-up, and normalization meet the specific needs of downstream applications.

Automated Schema Migration

This feature deserves its own shoutout, as 3rd-party APIs change schemas very often. Unless you have started investing in this feature, your data pipelines will break very often. Indeed, automated schema migration allows your ELT solution to automatically detect the schema of your data sources and map them to the appropriate target schema, thus avoiding a lot of errors and fixing.

Host and deploy your ELT

Once all your data connectors are ready, it's time to choose and prepare your hosting and provisioning strategy. This includes deciding whether to use on-premises infrastructure, or public cloud services, and how you want to deploy your compute - whether through virtual machines, containers, or managed services. You might want to look at a distributed architecture as data scales fast in volume.

Once your infrastructure is ready, you can proceed to deploy your in-house ELT solution.

Airbyte Open Source

Host and deploy Airbyte Open Source

If you choose to use Airbyte Open Source, the ELT part is already ready to use. Not only that, but you have access to a robust ELT tool trusted by 3,000 companies daily, syncing 900TB+ of data every month.

While you will also have to decide on and implement a hosting and provisioning strategy. Airbyte makes it easy to deploy its software. We offer the choice of installing it with a 1-command line script setup, Docker image, Kubernetes Helm chart, and Terraform template file that works with on-prem, public cloud, and containerized environments.

Secure Airbyte Open Source

Once your infrastructure is ready, you can proceed to deploy Airbyte Open Source on it. To ensure security, which is crucial when it comes to data, you will need to make sure that you secure Airbyte Open Source. The platform provides recommended best practices to follow, but you will also need to take into account your infrastructure's specific security requirements.

Build data connectors using Airbyte's CDK

With a catalog of 300+ data connectors to choose from; you will certainly find the ones you need. But for data sources that are not already supported, it can easily be done – in less than 30 minutes – by using Airbyte's low-code CDK. This process will require close to no technical expertise and just a basic understanding of the data sources involved. And when you build a data connector with Airbyte, you won't be the only one maintaining the connector; the Airbyte team, and community will!

Airbyte Cloud

If you're looking for a quick and easy way to set up an ELT platform for production use, Airbyte Cloud is the way to go. With just a few simple steps, you can create an account and start setting up your sources. The provisioning and deploying processes are managed by us; all you have to do is to set up your sources using our web UI. Similarly to Airbyte Open Source, you can use Airbyte's low-code CDK to build any missing data connector in under 30 minutes.

	Building in-house	Airbyte Open Source	Airbyte Cloud
From 0 to production	**$150,000**	**$33,000**	**$250**
Hosting, deployment and provisioning	10 days	3 days	
Secure the infrastructure hosting your ELT	depending on setup	depending on setup	
Build data connectors from scratch	8 days per connector	2 hours per missing conn.	2 hours per missing conn.
Incremental updates	10 days		
Change Data Capture support for databases	30 days		
Automated schema migration	60 days	30 days	
Idempotence (recovery from failure)	15 days		
Lossless replication	15 days		
Orchestration and automated scheduling	30 days		
Logic and process isolation	5 days		

Operate your ELT

This section will cover the key components of maintaining a reliable and effective ELT solution, including strategies for maximizing data integrity and availability, addressing technical issues and bugs, and upgrading and maintaining the solution over time.

DYI

Build or train an infrastructure team

Building or training an infrastructure team can be a significant undertaking. It's important to have individuals with the right technical skills and experience to manage the complexities of an ELT solution effectively. This could involve hiring new staff, training existing team members, handling departure, and potentially outsourcing to a third-party provider.

This team will need to ensure that the infrastructure hosting the ELT is reliable, secure, and scalable. This involves monitoring and maintaining the hardware, software, and networking components of the infrastructure, and handling any issues or outages that arise. The team must also be knowledgeable about the various tools and technologies used in the ELT solution, including data connectors, data processing frameworks, and workflow schedulers, and be able to troubleshoot these components on-the-fly as needed.

In addition to technical expertise, the infrastructure team must also have strong communication and collaboration skills, as they will need to work closely with other teams, such as data engineering, data science, and business operations, to ensure that the infrastructure is meeting the needs of these teams.

On-call shift

An on-call strategy is critical to ensuring the reliability and availability of your ELT solution. This involves having a designated team or individual who is available to respond to issues or incidents that may arise during and outside of regular business hours. In other words, someone should always be available to troubleshoot an issue. The strategy could include setting up alerts and notifications, defining escalation processes, and ensuring that the necessary tools and resources are in place to quickly diagnose and resolve issues. By having a robust on-call strategy in place, you can minimize downtime and keep your ELT solution running smoothly.

Factor in-house ELT and infrastructure downtime

Downtime can be a costly and disruptive issue for any organization – and it is even more critical when data is involved. It can lead to lost productivity, missed deadlines, and lost revenue. Therefore, it's important to understand the cost of downtime and develop strategies to minimize its impact. This could involve implementing redundancy and failover measures, having a solid disaster recovery plan in place, and proactively monitoring your ELT solution to identify and address potential issues before they cause significant downtime.

Monitoring

You can only fix something that you know is broken. Building a monitoring feature to ensure the successful operation of an ELT platform is essential. Monitoring is crucial to identify any issues in the data pipeline before they turn into major problems, which can ultimately lead to data loss and downtime. A basic monitoring dashboard should report metrics such as success rate average, sync status, and status per synchronization pipeline.

Logging

Once your monitoring tells you something is wrong, it's time to investigate! Logging is a critical tool for tracking and analyzing data movement, and it enables users to look into the status of their data pipelines in real time. A logging feature will offer valuable insights into how data is being processed, identify any potential bottlenecks, and troubleshoot issues.

Logging is also useful for auditing purposes, as it allows users to track the history of their data pipeline and maintain a record of all data movements. This feature can be crucial for data compliance purposes, as it ensures that users can maintain a clear and accurate record of data processing.

Capacity planning your infrastructure

Capacity planning is a critical aspect of maintaining the performance and reliability of your ELT solution. It involves forecasting your infrastructure needs and ensuring that you have the necessary resources in place to support your data integration requirements. This could include scaling up or down your infrastructure as needed, identifying potential bottlenecks, and implementing performance tuning and optimization strategies.

Fixing your ELT bugs

After the software engineering team builds the ELT solution, it is likely that a number of bugs will emerge over time, particularly as the system handles a growing number of use cases and increasingly large volumes of data.

Upgrading your ELT

Finding and fixing bugs in your ELT solution is only one aspect of the job. In addition, it is crucial to establish a well-defined upgrade process that minimizes downtime and ensures the ongoing reliability and effectiveness of your data movement.

Scaling your ELT

As your organization's data integration needs grow, it's essential to have a plan in place for scaling your ELT. This could involve scaling up your infrastructure, optimizing performance, or more advanced projects such as multi-cloud scale-out.

Troubleshoot your in-house data connectors

Because DIY connectors are usually bare-bone and not automated, they tend to require substantial maintenance after construction. Indeed, 80% of teams have to rebuild data connectors after deploying. When you start to have several in-house pipelines, this fixing pipelines' issues will become at least a full day a week.

Fixing pipelines could involve identifying and resolving data transformation or data loading issues, addressing data quality issues, and optimizing performance. Having a solid troubleshooting strategy in place is essential to ensuring that your ELT solution remains reliable and effective.

Airbyte Open Source

Although several elements mentioned in the DIY section are still relevant when using Airbyte Open Source, the platform offers a range of features and capabilities that can make the process of operating an ELT solution much easier.

Factor in infrastructure downtime

With Airbyte Open Source, the responsibility of handling potential bugs in the ELT is now the one of a team of data experts backed by over $180M of VC funding. The infrastructure team just needs to ensure that the hosting part is running smoothly.

Troubleshoot Airbyte data connectors

If you encounter any issues with Airbyte connectors, you can rest assured that you are not alone. With a vast community of over 600 contributors, 12,000 Slack members, and more than 30,000 companies using Airbyte, there is a higher likelihood of resolving any issues much more quickly than if you were dealing with your own custom data connectors. Unlike when maintaining your own connectors, the responsibility of maintaining Airbyte connectors falls on the Airbyte team and community, not just on the authors. This means that you can rely on a dedicated team of experts and an active community to help address any issues and ensure that your data integration processes are running smoothly.

Upgrading Airbyte

Although the process of upgrading Airbyte must be carried out manually, Airbyte Open Source provides an out-of-the-box mechanism for upgrading the platform while maintaining data integrity. Docker and Kubernetes deployments feature a one-command update to upgrade Airbyte to the latest version.

Monitoring

Airbyte Open Source provides basic monitoring and webhook failure notifications, to ensure that your data never stops flowing.

Airbyte Cloud

If you choose to use the Cloud version of Airbyte, you won't have to worry about managing the platform's infrastructure or ensuring its security and scalability. Airbyte takes care of all of these aspects, from hosting the solution to managing and securing it. Airbyte guarantees a 99% SLA support for certified connectors to ensure smooth and reliable performance. If any issues arise, the platform offers in-app chat support with an average response time of 10 minutes, so you can quickly resolve any problems and minimize downtime.

	Building in-house	Airbyte Open Source	Airbyte Cloud
Operate your ELT	**$423,000**	**$130,000**	**$10,000**
Troubleshoot your in-house data connectors	10 days per connector	3 days per connector	1 day per connector
Maintaining the ELT pipelines	12 days per connector		
Build or train an infrastructure team	depending on team	depending on team	
Capacity planning your infrastructure	20 days	10 days	
Handling on-call shifts	130 days	65 days	
Factor in-house ELT and infrastructure downtim	4 days	4 days	
Scaling your ELT	20 days	12 days	
Upgrading your ELT	9 days	4 days	
Logging	10 days	2 days	
Basic monitorign	10 days	2 days	

Grow your ELT

DIY

Continuously adding sources

As you start using new tools internally, those tools become new data silos that you will eventually want to include the data from, when you will be consolidating your data in the warehouse. The cost of having your engineers build and maintain a connector for every new data source, let alone add sophisticated new features and automations, will grow exponentially.

SSO

Single sign-on (SSO) is an essential security feature that allows users to access multiple applications and services with a single set of login credentials. SSO can help to reduce the risk of security breaches, streamline the login process, and make it easier to manage user access across multiple applications. Any good CSO will have SSO as a requirement. By implementing SSO for your ELT solution, you can improve the security of your data integration processes and simplify access for your users.

MFA

Multi-factor authentication (MFA) is another essential security feature that can help to protect your ELT solution against unauthorized access. MFA requires users to provide two or more forms of authentication before granting access to the system, such as a password and a security token. By implementing MFA, you can add an extra layer of security to your ELT solution and reduce the risk of security breaches and data theft.

Role-Based Access Controls

Role-based access controls (RBAC) are a critical feature of any ELT solution, as they allow you to control access to data based on the user's role within the organization. RBAC can help to ensure that only authorized users have access to sensitive data, reducing the risk of data breaches and unauthorized access. By implementing RBAC, you can maintain the security and integrity of your data integration processes and improve compliance with industry regulations and standards.

Audit Logging

Audit logging is an important feature that allows you to track and record user activity within your ELT solution. This can help you to monitor data access and usage, identify potential security threats, and improve compliance with industry regulations and standards. By implementing audit logging, you can maintain the security and integrity of your data integration processes and ensure that your data is being used in a responsible and ethical manner.

Workspace Management

Workspace management is a critical feature of any ELT solution, as it allows you to manage multiple projects and data integration workflows in a single environment. Workspace management can help you to streamline your data integration processes, improve collaboration among team members, and maintain the security and integrity of your data. This is especially true for large organizations that have multiple project and data teams. By implementing workspace management, you can improve the efficiency and effectiveness of your data integration processes and ensure that your team is working in a secure and compliant environment.

Secure your ELT

Because data is one of the company's most valuable but also sensitive assets, the engineering team will need to ensure that the solution is secure. Nowadays, security is not only about mitigating the risk of getting hacked, but also about being in compliance with data protection laws. A basic ELT security feature is to provide end-to-end data encryption when the data is in transit and at rest.

Data Compliance Features

Data compliance features are essential for ensuring that your ELT solution meets the necessary regulatory and legal requirements. This could include features such as data encryption, data masking, data anonymization, and data retention policies. By implementing these features, you can maintain the security and privacy of your data, reduce the risk of data breaches, and ensure that your data is being used in a responsible and ethical manner. Additionally, data compliance features can help you to comply with industry regulations and standards, reducing the risk of legal and financial penalties.

API access or CLI access

Building an API on top of your in-house ETL tool can significantly improve its usability and accessibility. More specifically, an API is required if you want to reduce the need for manual intervention and increase the speed of data processing by allowing programmatical data access and automating tasks. An API can also enable easier integration with other data tools along your data stack, such as data warehouses, business intelligence tools, and analytics platforms. Finally, an API can help to improve the security and reliability of your ETL pipeline by enforcing access controls and reducing the risk of errors or data loss.

A CLI can provide a quick and efficient way for users to interact with the ETL tool, allowing them to easily run tasks, check job status, and configure settings from the command line. This can be especially useful for the operation team, who will probably prefer to work from the terminal, as well as enabling the automation of tasks and integration with other systems. At times, a CLI can offer greater flexibility and customization options than a graphical user interface (GUI).

Integration with a scheduler

Building integration with schedulers such as Airflow, Dagster, and Prefect for your in-house ETL will be required to automate and manage complex data workflows, scheduling tasks and dependencies to run at specific times and ensuring that data pipelines are executed in a reliable and repeatable manner. This can be especially useful for organizations with large and complex data environments, where manual management of data workflows can be time-consuming and error-prone.

Allow advanced custom transformation

Sometimes, a specific data workload will require custom transformation that can only be done with a specialized tool such as dbt. Dbt is an open-source tool that enables users to transform data in a more structured and maintainable way, using SQL-based transformation logic that can be easily shared and collaborated on.

Security certifications

Depending on the industry your company is operating in or the type of data that are going through your data pipeline, passing SOC 2 or ISO 27001 certification for your in-house ETL tool might be a must. Additionally, passing SOC 2 or ISO 27001 certification can help organizations to build trust with their customers and partners, demonstrating their commitment to security and compliance and helping to differentiate themselves from competitors.

Airbyte Open Source

While most of the features that need to be built for a DIY solution will also need to be built for users picking Airbyte OSS, there are several key features that are handled natively by Airbyte OSS.

Allow advanced custom transformation

Airbyte Open Source integrated natively with dbt which allows users to create more sophisticated data transformations using SQL-based logic

Schema change detection

While Airbyte OSS is natively handling schema change, it will be missing a few advanced features that the administrator might want to have access to (automated behaviors).

API and CLI access, and integration with an orchestrator

API and CLI are natively provided by Airbyte OSS, providing users with greater flexibility and control over their data integration processes. Airbyte Open Source also offers out-of-the-box integration with popular schedulers such as Airflow, Dagster, and Prefect, enabling users to automate and manage complex data workflows in a reliable and repeatable manner.

Airbyte Cloud

By selecting Airbyte Cloud, you can be confident that you will always have access to the latest and greatest features, as our development team is continually working to improve the platform. Airbyte Cloud natively provides all the features mentioned above.

Our deep understanding of data integration, combined with ongoing communication with our users, allows us to prioritize and address the most commonly requested features. If you find that something is missing, we welcome your feedback and will work to incorporate it into our product roadmap. In fact, we provide visibility into our product roadmap, so you can see what features are currently being developed and when they are expected to be released.

	Building in-house	Airbyte Open Source	Airbyte Cloud
Scale your ELT	**$305,000**	**$197,000**	**$0**
SSO	20 days	20 days	
MFA	20 days	20 days	
Role-based Access Control	40 days	20 days	
Audit Logging	40 days	40 days	
Workspace Management	10 days	10 days	
Secure your ELT	15 days	7 days	
Data Compliance Features	20 days	20 days	
API or CLI Access	20 days		
Integration with a scheduler	5 days		
Allow advanced custom transformation	30 days		
Security Certifications (for regulated industries)	60 days	60 days	
PCI / HIPAA certifications with separation of data plane	30 days	15 days	

Total Costs

	Building in-house	Airbyte Open Source	Airbyte Cloud
TOTAL	**$879,800**	**$354,800**	**$22,250**
> From 0 to production	$150,000	$33,000	$250
> Operate your ELT	$423,000	$130,000	$10,000
> Scale your ELT	$305,000	$190,000	$0
Infrastructure / Airbyte Cloud costs	$1,800	$1,800	$12,000

Missed-opportunity cost

Choosing a ready-to-go ELT solution, such as Airbyte Cloud, over building one from scratch or using an open-source solution can save companies significant time and resources, and very importantly; avoid missed opportunities for business growth and innovation.

As we saw in this ebook, building an ELT solution (DIY) from scratch can be a complex and time-consuming process, requiring significant expertise and resources. And a DIY solution will often lack the ability to keep pace with the rest of the industry. A large majority (83%) of data leaders report feeling constrained by their current data integration setup, which limits their ability to adopt newer and more effective solutions, even when they are available.

On the other hand, open-source solutions can be attractive due to their lower cost, but they may not provide the same level of support, security, or functionality as a ready-to-go solution. Wakefield Research found that 70% of data leaders believe that their engineers could contribute more value to the business by concentrating on strategic projects rather than pipeline construction and maintenance.

By selecting a ready-to-go ELT solution, companies can quickly and easily access a robust, scalable, and reliable platform that is designed to meet the needs of their business. This can enable teams to focus on higher-value activities, such as data analysis, modeling, and innovation, rather than spending valuable time and resources building and maintaining an ELT solution. The missed-opportunity cost of not choosing a ready-to-go solution can be significant, as it can result in delayed decision-making, lost opportunities, and reduced competitiveness in the marketplace. Lost opportunities are a serious issue, 96% of data leaders whose organizations take days or longer to derive value from new data say that their companies would see improved business outcomes if data was able to move quickly.

While it is challenging to put a $ amount on it, we can expect the cost to be significant.

Conclusion

In summary, DIY is the least cost-effective solution for an ELT platform, costing a company $1,856,800, according to our findings. While Airbyte Open Source provides strong out-of-the-box features for ELT that cover basic use cases, companies with more sophisticated needs will need a strong engineering team to build additional features. Alternatively, Airbyte Cloud is the most cost-effective solution for ELT, scaling based on usage and covering both simple and advanced needs. In addition to the costs of building, operating, maintaining, and improving, missed opportunity costs are important to consider when choosing an ELT solution. With 57% of data leaders believing their engineers could build more advanced analysis if they weren't manually building data pipelines, and 81% of data leaders reporting that business decisions based on bad data have cost their company money, it is clear that trusting experts to take care of your ELT platform might be a wise choice.