The 7 best data pipeline tools of 2021 for eCommerce

The data pipeline is at the heart of your company's operations. It allows you to take control of your data and use it to generate revenue-driving insights.

However, managing all of the operations involved in your data pipeline (data extractions, transformations, loading into databases, orchestration, monitoring, and more) can be a little daunting.

Here, we present the 7 best data pipeline tools of 2021, all of which can help you to take control of your data pipeline:

1. Free and open-source software (FOSS)

Free and open-source tools (FOSS for short) are on the rise. Companies opt for FOSS software for their data pipelines because of its transparent and open codebase, as well as the fact that there are no costs for using the tools.

Among the most notable FOSS solutions are:

- petl, Bonobo or the Python standard library - software that helps you to extract data from its sources.
- pandas - with its Excel-like tabular approach, pandas is one of the best and easiest solutions for manipulating and transforming your data, just like you would in a spreadsheet.
- Apache Airflow - a cron job on steroids. Apache Airflows allows you to schedule, orchestrate, and monitor the execution of your entire data pipeline.
- Postgres - one of the most popular SQL databases. Postgres adds to the usual feature set of SQL databases by extending its data type support (covers unstructured data with JSON fields) and offering built-in functions which speed up analytics.
- Metabase - a lightweight application layer on top of your SQL database, which speeds up querying and automates report generation for the non-technical user.

PROS:

- **Free**. There are no vendor costs.
- **Fully customizable**. Open source means that you can inspect the code and see what it does on a granular level, then tailor it to suit your specific use case.
- **No vendor lock-in**. No contractual obligation to keep with a vendor who doesn't fulfill your needs.
- **Community support**. FOSS has a community of fans who offer plenty of support on StackOverflow and other channels.
- **Fun**. FOSS solutions allow for a lot of tinkering, which - we're ready to admit it - is fun.

CONS:

- **Solution lock-in**. Customized solutions are hard to disentangle when moving to a different tool or platform, especially when home-brewed solutions do not follow the best engineering practices.
- **High maintenance costs**. Every change to the data pipeline requires you to invest engineering hours… and data pipelines change a lot. From APIs altering their endpoints to software upgrades deprecating libraries, FOSS solutions are guilty of high maintenance costs.
- **Lack of technical support**. When things go wrong, there's no one to call who can help you resolve your technical mess. You must be more self-reliant and budget for errors.
- **Scaling**. As your company grows, so do your needs. The engineering solutions differ drastically depending on the scale of your data operations. For example, implementing the infrastructure for a distributed message broker makes sense when you are processing high volumes of streaming data, but not when you are collecting marketing spend via APIs. FOSS solutions require you to develop in-house expertise in scaling infrastructure (costly) or outsource it to contractors instead (also costly).
- **Time-to-insights opportunity costs**. The average time it takes to build your entire data pipeline is north of 9 months. Vendor solutions shorten the timeline from months to weeks, so you skip the opportunity costs accumulated when waiting for your BI infrastructure to be ready to answer questions.

Who is it for?

- Data-scarce companies who do *not* plan to scale.
- Small data pipelines, which are developed as prototypes within a larger ecosystem.
- Hobbyists and tinkerers.

2. Keboola

Keboola is a Software as a Service (SaaS) data operations platform, which covers the entire data pipeline operational cycle. From ETL jobs (extract-transform-load) to orchestration and monitoring, Keboola provides a holistic platform for data management. The architecture is designed modularly as plug-and-play, allowing for greater customization. In addition to all of the expected features, Keboola surprises with its advanced take on the data pipeline, offering one-click deployments of digital sandboxes, machine learning out-of-the-box features and more.

PROS:

- **State-of-the-art architecture**. The engineering behind Keboola is superb. It is resilient, scales effortlessly along with your data needs and uses advanced security techniques to keep your data safe.
- **Fully customizable**. Keboola supports a variety of scripting languages (SQL, Python, R), giving you the option to fully customize the data pipeline on your own language terms. Additionally, it builds on top of open-source technologies such as Singer, which

allows you to create your own features (e.g. add an obscure data source that is not already covered by their extractors).

- **Audit & security**. The platform offers an exhaustively auditable data pipeline by versioning and fingerprinting changes. This is just one of the solutions built on top of the existing advanced security measures, leveraging AWS's best-in-class security standards.
- **Beyond ETL**. Keboola boasts a suite of transformative technologies built on top of the ETL: scaffolds to deploy end-to-end pipelines in just a couple of clicks, data catalogs which allow you to share data between departments (breaking those silos) and document data definitions, and digital sandboxes that allow for experimentation and prototyping without affecting the underlying engineering.
- **No vendor lock-in**. Monthly fees keep your relationship with Keboola flexible. Unlike the majority of vendors, it is easy to take your data and scripts out of Keboola and migrate them to a different solution.
- **Freemium track**. No sales talk before implementation. Simply give it a spin with the free forever plan. It's more than enough to get you started.
- **More than 1200 components** - extractors & writers. There's a component for your use case. And even if there isn't you can simply build it or re-use generic extractor and generic writer components.

CONS:

- No database replication based on changelogs, but does offer automated table snapshot, backup, and recovery.
- Implementation requires a bit of technical know-how.

Who is it for?

- Fast-growing startups and companies that are scaling rapidly.
- Medium-sized companies who are looking for same-day data delivery and real-time data insights.
- Enterprises and big data deployments looking for an easy-to-manage, all-in-one solution for their data pipeline.

3. Stitch

Stitch is an ETL platform which helps you to connect your sources (incoming data) to your destinations (databases, storages and data warehouses). It is designed to enhance your current system by smoothing out the edges of ETL processes on data pipelines.

PROS:

- Stitch has one of the most extensive integrations of all vendors. It covers a vast range of sources and destinations.
- Relies on the Singer framework, which allows you to customize parts of the pipeline yourself.
- It offers cron job-like orchestration, as well as logging and monitoring. This allows you to keep an eye on the health of your data pipeline.
- Stitch offers a free trial version and a freemium plan, so you can try the platform yourself before committing.

CONS:

- A lot of integrations (sources and destinations) require a higher payment plan, meaning that your scaling may be hindered by steeper costs.
- No automated table snapshot, backup, or recovery. If there is an outage or something goes wrong, you could suffer data loss.
- Limited transformation functionalities. Unlike its sources and destination integrations, Stitch is lacking when it comes to transformation support. It also requires additional staging storage to compute data transformations.
- It does not offer 24/7 live support.

Who is it for?
Companies who prefer a synching data pipeline with a lot of integrations (Stitch offers a high number of integrated sources and destinations), but have low requirements for transformations and do not plan to scale horizontally to new integrations.

4. Segment

Segment is a customer data platform which helps you to unify your customer information across your technological touchpoints. With its clickable user-interface, Segment offers an easy-to-use platform for managing integrations between sources and destinations. Its platform is centered around users; all of the data transformations, enrichment and aggregations are executed while keeping the user at the center of the equation.

PROS:

- Identity stitching. One of the major advantages of Segment is that it offers identity stitching. It uses an identity graph, where information about a customer's behavior and identity can be combined across many different platforms (e.g. Google, Facebook...) and clients (e.g. desktop, phone…). This enables you to centralize customer information.
- Personas. Segment automatically builds up personas based on your data. Personas can be used to streamline marketing and sales operations, increase personalization, and just nail that customer journey in general!

CONS:

- Price. Segment does have a free tier, but it's unusable for anyone who has more than two data sources. Many of its worthwhile features are locked behind higher-tiered plans, and customers complain about how expensive it has become (a lot).
- Non-user based analytics. Segment has devoted a lot of its development to user analytics. If your needs exceed those of customer-centric analyses (e.g. revenue reports, internet of things, etc.) Segment might not offer the best support for your use case.

Who is it for?
Segment is ideal for companies who would benefit massively from stitching their customer information across platforms (and have the budget to do so).

5. Fivetran

Fivetran is an ETL platform which technically automates ETL jobs. It enables you to connect your data sources to your destinations through data mappings. It supports an extensive list of incoming data sources, as well as data warehouses (but not data lakes).

PROS:
- Extensive security measures make your data pipeline safe from prying eyes.
- Supports event data flow, which is great for streaming services and unstructured data pipelines.
- It allows you to access the data pipeline with custom code (Python, Java, C#, Go…), thus making it possible to build your connections.

CONS:
- Limited data sharing options.
- No open source. Fivetran does not showcase (parts of) its codebase as open-source, making it more difficult to self-customize.
- Vendor lock-in. Annual contracts make it harder to separate yourself from Fivetran. In addition, it's currently impossible to take your data, schemas and queries and easily migrate them to another platform.
- Limited to non-existent data transformation support. It does not transform data before loading it into the database, but you can transform it afterwards using SQL commands.
- Requires additional staging storage to compute data transformations.

Who is it for?
Fivetran is geared more towards data engineers, analysts and technical professionals. It is great for companies who plan to deploy the tool among their technical users, but not for those who want to democratize data pipelines across the board.

6. Xplenty

Xplenty is a data integration platform which connects your sources to your destinations. Through its graphical interfaces, users can drag-and-drop-and-click data pipelines together with ease.

PROS:

- The visual editor is intuitive and fast, making data pipeline design easy. This also allows non-technical users to access data pipelines and collaborate across departments.
- It does not require coding ability to use the default configuration.

CONS:

- No ability to add your own data source.
- Limited data sharing options.
- Vendor lock-in. Annual contracts make it harder to separate yourself from Xplenty.
- Limited logging and monitoring. Not all logs are available and it is hard to inspect the platform when things go wrong.
- It does not offer as many 3rd party connectors as other platforms.
- Lacks real-time data synchronization
- No on-premise solution.

Who is it for?
Companies who are looking for a cloud-based solution which is easy to use, but does not require a lot of modifications or scaling.

7. Etleap

With its clickable user interface, Etleap allows analysts to create their own data pipelines from the comfort of the user interface (UI). Though sometimes clunky, the UI offers a wide range of customization without the need to code.

PROS:

- Strong security standards keep your data safe.
- No need to code in order to use the transformation features.
- Covers a wide variety of incoming source types, such as event streams, files, databases, etc.

CONS:

- Limited destinations - Amazon Redshift, S3 Data Lakes, and Snowflake only.
- No REST API connector.
- The user interface is not as friendly.

Who is it for?
Analysts and data engineers who want to speed up their data pipeline deployment *without* sacrificing the technical rigor to do so. Not so apt for non-technical users, since it requires an understanding of underlying engineering standards to use the platform.

Which tool should you choose?

In short, it all depends on your use case:

Data pipeline tool	Best for
Keboola	● Fast-growing startups and companies who are scaling rapidly. ● Medium-sized companies who are looking for same-day data delivery and real-time data insights. ● Enterprises and big data deployments seeking an easy-to-manage, all-in-one solution for their data pipeline.
Free and open-source software (FOSS)	● Data-scarce companies who do not plan to scale. ● Small data pipelines, which are developed as prototypes within a larger ecosystem. ● Hobbyists and tinkerers.
Stitch	● Companies who prefer a synching data pipeline with a lot of integrations, but have low requirements for transformations and do not plan to scale horizontally to new integrations.
Segment	● Companies who would benefit massively from stitching their customer information across platforms (and have the budget to do so).
Fivetran	● Data engineers, analysts, and technical professionals. It is great for companies who plan to deploy the tool among their technical users, but not for those who want to democratize data pipelines across the board.
Xplenty	● Companies who are looking for a cloud-based solution which is easy to use, but does not require a lot of modifications or scaling.
Etleap	● Analysts and data engineers who want to speed up their data pipeline deployment *without* sacrificing the technical rigor to do so. Not so apt for non-technical users, since it requires an

	understanding of underlying engineering standards to use the platform.