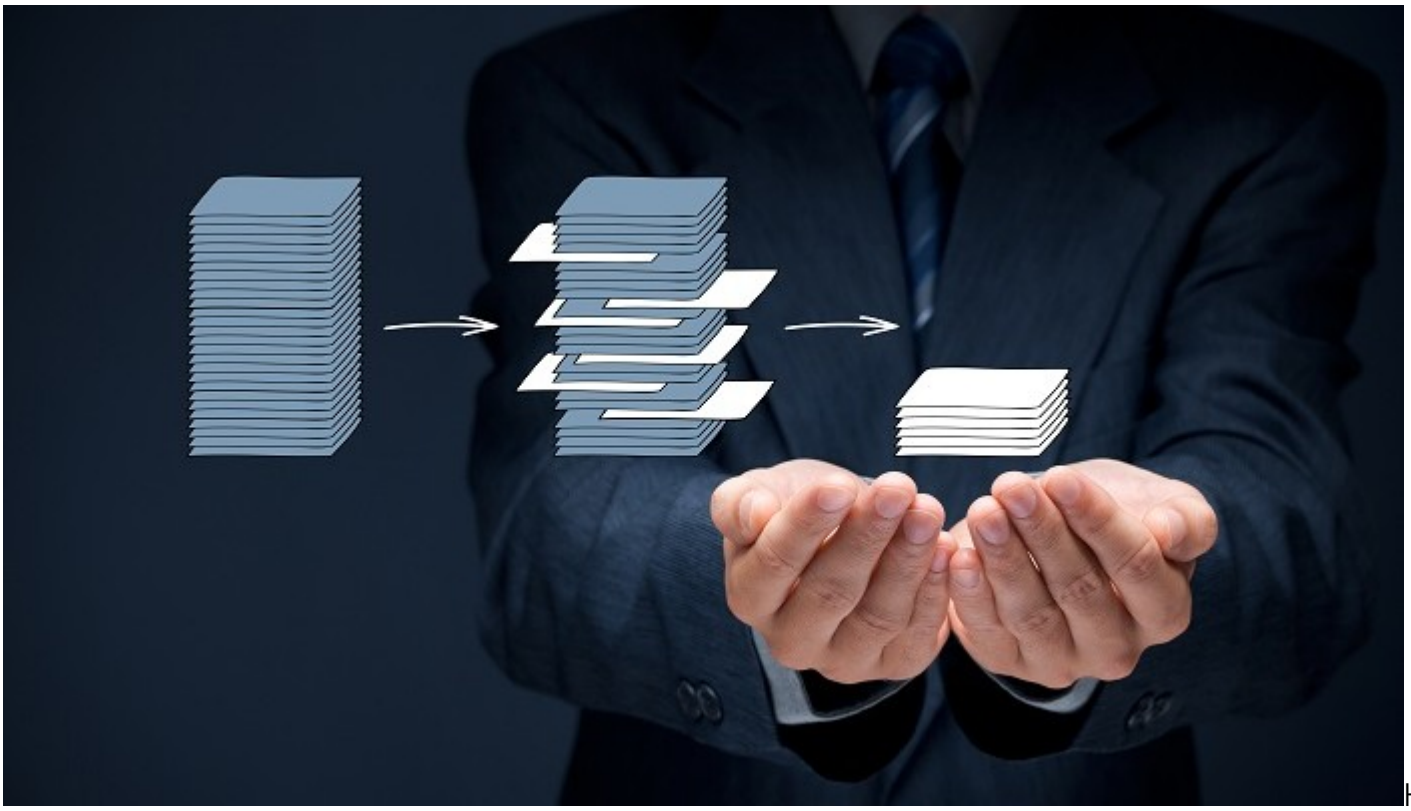
4 REASONS TO AUTOMATE DATA INGESTION





Hardly a day goes by without talk of automation and big data in any company. These days, the market understands the need for data: it's the de facto way to gain business intelligence. And, data science

and machine learning are go-to tools in predictive analytics, which means you need data, and a lot of it.

But data must be cleaned and ready to go, in formats that allow for data analysis. This process, known as data ingestion, is something you should be automating.

What is data ingestion?

Data ingestion refers to the ways you may obtain and import data, whether for immediate use or data storage. Importing the data also includes the process of preparing data for analysis. In a broader sense, data ingestion can be understood as a directed dataflow between two or more systems that result in a smooth, and independent, operation (a definition which already implies some independence or automation).

Ingestion can occur in real-time, as soon as the source produces it, or in batches, when data is input in specific chunks at set periods. Generally, three steps occur within data ingestion:

1. **Data extraction:** Retrieving data from sources
2. **Data transformation:** Validating, cleaning, and normalizing data to ensure accuracy and reliability (sometimes known as trustworthiness)
3. **Data loading:** Routing or placing the data in its correct silo or database for analysis

Of course, as data grows, this three-step process gets bigger and takes more time. Historically, data ingestion was manual, relying on manual data gathering and manual importing into a custom-built spreadsheet or database. In that process you may correct for data inaccuracies to ensure the data is similar, but human error can't ensure 100% clean or trustworthy data.

Today, in the age of big data, manual data ingestion is rarely possible. Companies have numerous data sources, often totaling hundreds, with data coming in 24-hours a day. And the really fun part? Data inputs as a variety of formats, so companies need to convert the data to similar formats. More and more, companies are implementing automation in order to ingest data for efficient data analysis.

Reasons to automate data ingestion

Reasons to automate are countless and will vary from company to company. Still, the biggest takeaways from data automation are quite clear. Here's what automating data ingestion will do for you:

1. Improve time-to-market goals

In 2016, 55% of B2B companies say their inability to merge data from a range of sources in a timely way actually holds them back from achieving these goals. This makes sense, when analytics projects often take three times longer than people expect. Frequently, companies spend time preparing for the analysis, but if the data ingestion and data preparation hasn't gone smoothly or efficiently enough, there's no data to analyze, delaying initial goals. And if you can't get your product to market in time, you've lost your competitive edge.

2. Increase scalability

Stepping into the world of automatic data ingestion may feel overwhelming, especially trying to adapt data science and machine learning techniques. The good news is that it's easy to remain small as you're automating – pick one or two data sources and work out the best way to automate, relying on industry best practices. As you grow more comfortable and free up time, you can scale up, automating more data over time.

As you automate more, automating becomes easier, especially with the implementation of self-service tools. As new data sources are identified, a centralized IT group doesn't have to field and implement every single request for a data source, if there's a self-serve, automation tool that can help establish a data source.

This scalability is particularly beneficial when part of the infrastructure or service requirements change – which is inevitable. While an automated ingestion process may require some manual tweaking, you won't have to waste valuable time and money retraining a team on how to alter ingestion techniques. Instead, the operation is smooth and sees significantly less interruption.

3. Refocus on necessary work

Preparation is key in any project but imagine spending 4/5 of your time on tedious tasks before doing work that produces results. Data scientists repeatedly report that the least interesting, desirable, or challenging part of their work is the data preparation – that part of data ingestion that gets the data ready for analysis.

Statistics indicate that up to 80% of an analytics project is dedicated to this task, not to the wider challenge of applying or developing particular algorithms and analyzing the results. Instead, your expert data team is busy with tedious tasks like extracting data from various apps, transforming formats with custom code, and loading data into various siloed systems.

By automating the system, your data scientists are freed up to perform the work both they and the company wants: analysis, leading to marketable changes and improvements.

4. Mitigate risk

We understand that data is key in business intelligence and strategy. Without it, today, you'll be quickly pushed aside by companies with sharper completive edges. That's a risk you can't afford to take.

Automating data also mitigates other risks: the risk of human error in extracting, transforming, and loading data. The risk of falling behind because you can't keep up with the data you are collecting. (This can lead to extreme situations where the only way to catch up is to actually let go of data altogether – a significant waste of resources.) The risk that your company could be doing more,

Data efficiency is the goal

The bottom line to all of this is that automating data ingestion is more efficient, which saves time and money. The more scalable you are, the easier it is to bring more data into the fold, without risking your time-to-market goals. Automating data ingestion promotes increased scalability and efficiency.

Need one more reason for data ingestion? Your data department will like you a lot more! By

reducing the need for tedious tasks, employees can focus on the work they prefer – the challenge of and the information that comes from data.

Additional resources

For related reading, explore these resources:

- BMC Machine Learning & Big Data Blog
- What Is a Data Pipeline?
- Enabling the Citizen Data Scientists
- Structured vs Unstructured Data: A Shift in Privacy
- Data Management vs Data Governance: Main differences