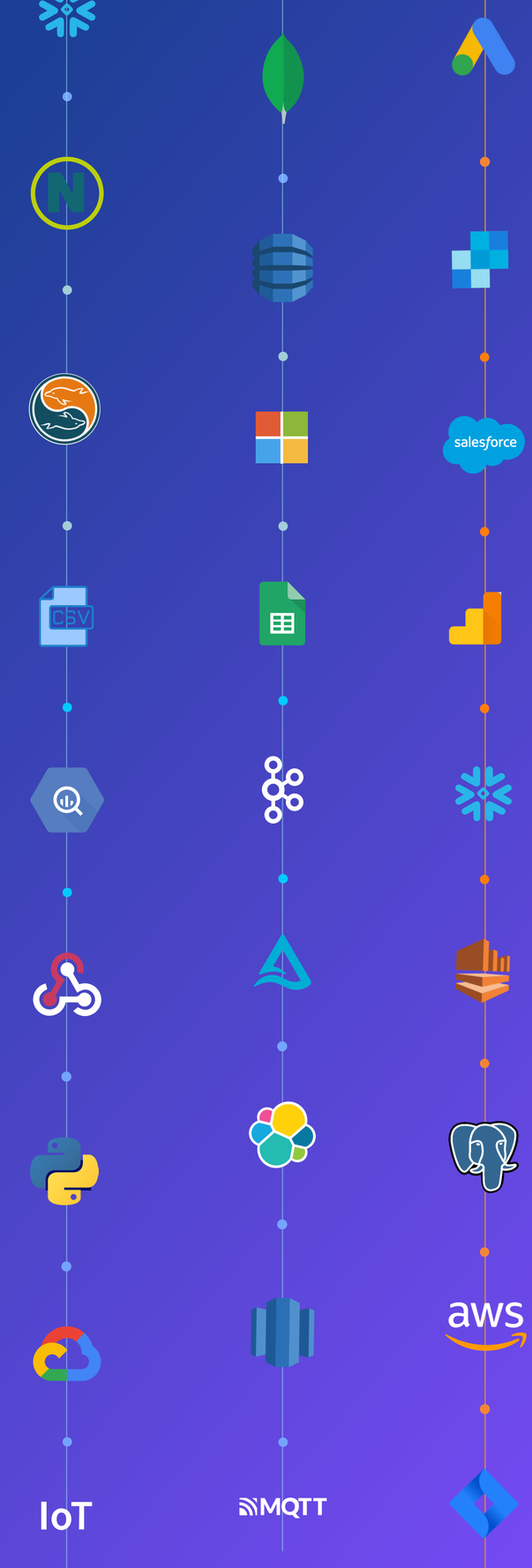
# Top data integration qualities to watch out for in 2023

InComparison Paper by **Bloor**
Author **Daniel Howard**



# Introduction

**Data integration is a set of capabilities that allows data  that is in one place to be moved to another place (although for clarification's sake, we should mention that – unlike, say, the Windows move operation – this does not generally remove or alter the original data). This is more complicated than it sounds. For instance, owing to different requirements in your source and target systems, the metadata surrounding your data will often need to change even as the data itself remains essentially the same. Hence you will often have to both physically copy your data to the target system and transform its metadata into a form suitable for its new location.**

This is usually accomplished by applying a set of transformations to your data, and its associated metadata, during the data integration process. ETL (Extract, Transform, and Load) is the most widely known of this type of methodology (in fact, ETL is probably better known – at least as a term – than data integration itself). Lesser-known alternatives include ELT (Extract, Load, and Transform) and ETLT (Extract, Transform, Load, and Transform again), with the major difference between these methodologies being – as you might suspect from their names – at what point (or points) in the process the transformations are applied. This has various architectural, performance, and security implications, but suffice it to say that all three have their place in data integration.


> The metadata surrounding your data will often need to change even as the data itself remains essentially the same.

**Other types of data integration practices include data replication, CDC (Change Data Capture) and associated techniques, which essentially just copy data without transforming it (but with other benefits instead). For example, CDC is frequently used for monitoring data sources for new or changed data, capturing those changes, then forwarding them elsewhere for processing and enabling change propagation. This allows you to keep downstream systems in sync with your data sources in (near-)real-time. Of course, it is also possible to combine CDC with ETL (or ELT or ETLT) if transformations are required as part of this process (or vice versa).**

There are multiple use cases for data integration. Two of the most notable are to support data migrations – of which migrations to the cloud are particularly prominent at the moment – and to move data from an operational database to a data warehouse in support of data analytics. In the latter case, data warehouse automation tools extend this capability by understanding the relevant schemas and helping to automate the creation of said warehouses. Moreover, they typically use replication and/or CDC in order to ensure that the target system is kept up to date, in a good example of what we were talking about at the end of the previous paragraph. Note that for analytics in particular, data virtualisation provides a viable alternative to data integration by allowing you to query data that exists in multiple physical locations as if it sat within a single data warehouse. This means that you do not need to physically move it, which tends to greatly simplify things. Of course, data virtualisation cannot be applied to data migration and similar use cases, where the movement of data is not an incidental difficulty but rather the entire point, making it somewhat niche despite its advantages.

Data integration has also changed and evolved substantially in response to recent data trends and market factors. The most prominent of these trends is the cloud, which has created a growing demand for data integration technologies that can effectively facilitate cloud migrations and has made cloud compatibility an expected capability across more or less every data space. AI and machine learning have also remained popular, and data integration tools are frequently called on to migrate training data (often repeatedly) in order to train accurate machine learning models. This has resulted in some data integration solutions developing capabilities that speak to this use case in particular, such as model drift detection. There is also growing interest in leveraging data integration and its adjacent technologies on data that sits in the mainframe (in order to run analytical queries, for example) in combination with data that doesn't. Since the ideal situation in this scenario is that mainframe data doesn't leave the mainframe, the adoption of data virtualisation as a supplementary capability to data integration has been a recent trend.All that said, the insight at the heart of this document is that the vast majority of data integration solutions – and certainly all the data integration solutions worth taking seriously at the enterprise level – have several essential capabilities in common, that on the one hand are "must-haves" for a data integration solution to function effectively, but on the other are so widespread among such solutions that they are effectively table stakes. Examples of this would be things like cloud integration, a reasonably robust suite of connectors to various data sources, a graphical user interface, some fairly large set of pre-built transformations for use with ETL and/or ELT, and so on. Our aim in this document is to discuss the five most important qualities that a data integration solution should strive for beyond these basic capabilities. This includes what those qualities are, how they can be achieved in the context of data integration, and how they help generate value from data integration.

**Data virtualisation provides a viable alternative to data integration by allowing you to query data that exists in multiple physical locations.**

# Self Service

**There is an increasing trend towards the provision of self-service capabilities within (and, for that matter, without) the data integration space. This often (though not always) implies the use of visual, no- or low-code, drag and drop development environments to build your transformations and data pipelines, such that building the aforementioned pipelines and transformations is accessible to both technical and nontechnical users. Some products have separate interfaces that cater to technical and non-technical users separately, but this is often not ideal as it can have deleterious effects on collaboration. Other features that promote ease of use, such as out of the box accelerators, large and robust sets of prebuilt transformations, and so on, can also contribute to a greater degree of self-service.**

The advantage of self-service in regards to data integration is that it allows your data consumers to move the data they need to wherever they need, whenever they need it, without needing to go through IT to do so. This allows both your data consumers and IT to be more efficient, since the former can go ahead and do whatever it is they need to do without having to wait on a who-knows-how-long process, while the latter doesn't need to get involved in who-knows-how-many requests for data movement. Moreover, as far as analytics use cases are concerned, the window of opportunity for analysing the data and getting useful results can be quite tight. Self-service data integration systems allow you to move data more expediently, analyse it sooner, and take advantage of more of those windows of opportunity.

> There is an increasing trend towards the provison of self-service capabilities within the data integration space.

# Collaboration

**Collaboration serves as a natural step onward from self-service. In many spaces, this is referred to as the democratisation of data.**

It can be achieved through both explicit collaboration mechanisms, such as data sharing and workflows, or through looser techniques, such as by enabling users to leave comments and/or ratings on any given data asset. The goal is that in addition to enabling more expedient collaboration when such things are directly necessary, you are also enabled to capture, centralise and expose in-house "tribal" knowledge about your data assets that would otherwise be strewn across many of your individual users. This allows you to benefit from the domain expertise that resides in your organisation in a centralised and systemic way. Moreover, this sort of centralisation means that expertise is not lost when, for example, a long-standing domain expert decides to retire. Data integration solutions can equally benefit from this (data pipelines are still a kind of data asset, after all, and can very well be collaborated upon) and facilitate it, say, by enabling automated workflows driven by data pipelines.

"Collaboration serves as a natural step onward from self-service.

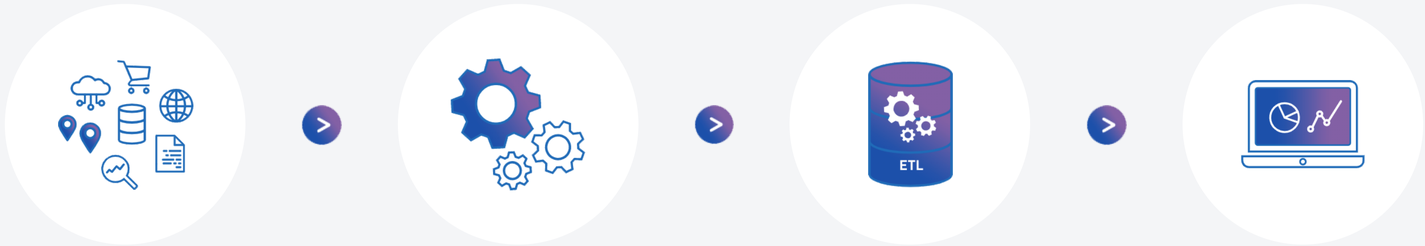
# Automation

**Automation has always been a perennial concern for data integration: it makes life easier, supports self-service and collaboration, reduces costs, and improves efficiency.**

Today, there is some emphasis on using AI and machine learning to drive automation, and while it is certainly true that these technologies can introduce automation, in a variety of ways no less, this does not mean that machine learning or AI are necessary to deliver automation. As an example, machine learning can be used to drive recommendations, but is not necessary for it. Moreover, to reap the benefits of machine learning you will need accurate, well-trained models, and for that you will need a robust pipeline of training data, that is reusable in order to combat data drift.

Data integration is well-suited for this, making it both a driver and recipient of automation. Other automation capabilities, perhaps more mundane but no less useful, include such things as automated workflows and data pipelines, which are often instrumental for fully achieving the benefits of self-service and collaboration that we have already mentioned.

Automation as a whole is self-evidently important, and the automated capabilities present in a data integration product can prove a particularly significant differentiator. The most highly automated solutions will feature a variety of embedded automation and machine learning throughout their built-in data processes and may even provide extensible automation capabilities as well.




> **The most highly automated solutions will feature a variety of embedded automation and machine learning throughout their built-in data processes.**

# Innovation

**Innovation is an inherently more nebulous quality than the others we have discussed here, but it should be obvious why it's desirable: companies and products that innovate are necessarily doing things in new ways that are entirely (or at least mostly) unique to them.**

If you want a leg up on your competition, investing in a solution that innovates, or intends to innovate, is a good start. Of course, the risk you run is that innovation doesn't always pan out well, meaning that vendors with a history of innovating repeatedly and effectively should be regarded very highly indeed.


Moreover, data integration is a mature space that is closer to being stagnant than it is to being volatile. At the same time, innovation is necessary to take full advantage of modern data trends and maximise the benefits thereof, even – or perhaps especially – in a mature space like data integration. As a result, being able to demonstrate both a willingness to innovate and an aptitude for doing so is a very appealing trait for a data integration solution to have.

> **Data integration will continue to be an essential part of business innovations and this will help businesses improve the whole experience for the consumer.**

# Acceleration

**Finally, all of the above qualities should come together to create a solution that accelerates your data integration processes to the greatest extent possible. This isn't just about automating data pipelines, or enabling self-service, or what have you, but combining everything into a cohesive whole that dramatically increases the overall speed and efficiency at which you can move and transform your data, and thus allows you to operate on and derive value from that data more quickly, more readily, and at greater volume.**

In terms of what a data integration solution can do to enable acceleration specifically, it's usually a matter of delivering the qualities and capabilities we have already discussed in a way that is at least as great, if not greater, than the sum of its parts, with the key being that automation, collaboration and self-service features need to feed into each other and integrate fluidly, and be leveraged either in concert or in sequence, not just individually.

There is also the matter of performance and scalability: you will want your data integration solution to perform well at small loads and scale out to perform competitively at large ones. Cloud integration can be particularly useful here, owing to the dynamic/elastic scaling that the cloud can provide.

> **Combining everything into a cohesive whole that dramatically increases the overall speed & efficiency at which you can move and transform your data.**

# Conclusion

**Data integration is a mature space, and many capabilities within it have become little more than rote for the vast majority of data integration solutions.**

It is all too easy to simply dismiss the space as stagnant, and rest on your laurels with a functional but ultimately archaic and inefficient data integration system.  It is our hope that this eBook has exhorted you to consider your choice of solution more carefully, in light of the five particularly valuable qualities we've discussed, and perhaps even to examine some of the more recent entries into the space that you might otherwise have overlooked. Most importantly, we hope to have provided a set of criteria that is relevant for assessing and comparing data integration solutions within the modern data landscape, in greater detail (and providing greater differentiation) than anything that could be boiled down to a checkbox list of table stakes features.

Bloor

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

Tel: +44 (0)1494 291 992
Web: www.bloorresearch.com
Email: info@bloorresearch.com

**For further reading on this subject, please visit the following articles / weblinks on the Bloor website**

www.linkylinkylinky.com
www.linkylinkylinky.com
www.linkylinkylinky.com