MarketUpdate

Stream Processing

Market Basics

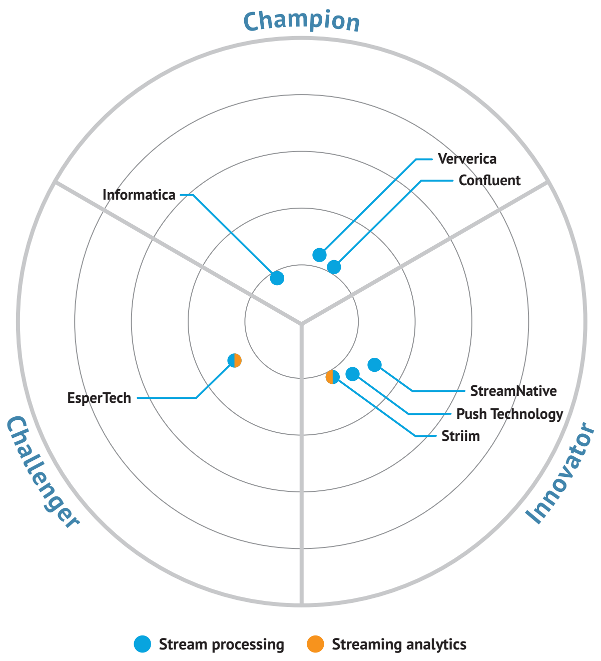
Stream processing solutions, broadly speaking, exist to ingest, move and/or transform streaming data. Streaming data, in turn, is data that is generated (and hence must be processed) continuously from one source or another. The core idea driving the space is that you will often benefit from being able to ingest, process, and act on your data in real-time and on a continuous basis, accomplished via stream processing, rather than with a significant time lapse, as in traditional batch solutions. This is particularly true for data that is highly volatile, meaning that it becomes stale very quickly, and if you do not act on it immediately you may not get to act on it (productively) at all. Even with more stable data, being able to utilise it more quickly is frequently a very good thing. Streaming technology also offers a one-to-many approach to real-time data: unlike simple messaging, stream processing allows any number of applications to ingest and react to the same streamed data at the same time, and is therefore significantly more powerful and useful.

The space has always had a particularly notable open-source presence, to the point that we consider it not just a major trend, but effectively foundational for the space. Apache projects such as Flink, Spark, Pulsar, and Kafka have generated a lot of attention for and within the streaming space, and they remain popular despite the growth of competing proprietary solutions. That said, open-source streaming projects tend to be narrower in scope than their proprietary counterparts, and although it is entirely possible to build an open-source streaming solution, it will largely involve assembling it yourself from several different open-source offerings. You will need solutions for data flow management, distributed messaging, and stream processing itself, at a bare minimum. You will also end up without ongoing enterprise support unless you subscribe to one (or more) of the vendors that provide such, but that instead removes one of the key draws for an entirely open-source stack. On the other hand, the technology itself is often highly competitive, and the community-driven approach of open-source is clearly appealing.

It is fortunate, then, that any open-source streaming engine worth talking about has at least one, if not more, proprietary offerings built using it. The makeup of these offerings is varied: they range from essentially just the technology itself, but with enterprise support added in, to full-blown holistic streaming platforms built off of the underlying open-source technology. Of course, there are also a range of proprietary solutions that use closed-source streaming technology to the same ends. Regardless, in all cases they demand far less effort on your part than a home-grown solution while frequently offering far greater functionality (such as the ability to intelligently mask or otherwise transform streaming data on the fly), and it is not difficult to see why someone might (and, more often than not, should) prefer it over building their own.

We should also take a moment to mention streaming analytics. Streaming analytics is a space that is closely related to – but ultimately distinct from – stream processing. It exists to extract actionable insights from your streaming data, usually as it enters your system, and is sometimes (though hardly always) offered in a joint capacity with stream processing. We discuss this topic further in our recent Market Update.

Figure 1:
The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator or Challenger segments, depending on their innovation score. The exact position in each segment is calculated based on their combined innovation and overall score. It is important to note that colour coded products have been scored relative to other products with the same colour coding.



Market Trends

The streaming space in general has grown significantly over the past few years, and we have every reason to think this trend will continue. Several factors, such as the increasing popularity of the cloud, the Internet of Things (IoT), the widespread implementation of 5G, have created a significant increase in the amount of streaming data that is available to most organisations, thus driving the adoption of streaming technologies. And with exponentially more streaming data (as well as more data in general) coming in every year, there is a definite need for highly performant, highly automated streaming solutions that are well-suited for handling this increased throughput. There is also a greater appetite for real-time technology in general, perhaps brought on – at least in part – by the fact that customers increasingly expect immediacy from their applications. Where a slow-to-update app would certainly have been frowned on a few years ago, these days it could easily mean a lost customer (or rather more to the point, many lost customers). Moreover, stream processing's presence can now be felt across a wide range of industry verticals, where previously only a few had really taken to it. In short, streaming technology has gone from burgeoning – but still essentially niche – to mainstream.

We have identified several discrete trends within the stream processing space. As such, we have divided further discussion into sections for ease of consumption.

General data management trends

The increasing popularity of the cloud, of containers, and of IoT (among other things) is impacting almost every space within data management. Streaming is by no means an exception. IoT, for instance, has always been a driver for the space, and its greater prevalence has served to further drive demand for streaming technologies. As you might expect, this is particularly true for the kind of highly performant and scalable streaming solution that can readily handle processing sensor data that is arriving at a massive scale.

Change Data Capture, or CDC, has also emerged as a driver for stream processing, in the sense that an appropriate stream processing solution can, for example, enable you to stream changes in your data to your other systems in real-time. This has a clear application for monitoring and analytics use cases (such as fraud or anomaly detection), particularly when they are backed up by AI models, in that it allows you to see how your data is changing, and analyse those changes, as they happen, rather than some time after the fact. Imagine the difference between detecting a fraud attempt as it is ongoing vs. uncovering it the next day, for instance.

Deployment to a range of clouds is widely supported within the space, although this is hardly new. It has spurred some vendors to actively target (or, more accurately, continue to target) customers that are undergoing cloud migrations, by effectively delivering a combination of data integration and stream processing features. This allows said vendors to address both an initial batch migration and ongoing streaming ingestion after-the-fact. Moreover, all three of the major cloud providers – Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) – offer their own streaming solutions within their respective platforms. We discuss these solutions in more detail in the next section, but as far as the market is concerned their major impact is that they have introduced a lot of cloud users to streaming for the first time, either as a solution in themselves or as a jumping off point. In this sense, the cloud has been a very substantial driver for the streaming space. It has also had the effect of normalising features and pricing structures that are particularly conducive to the cloud: dynamic scaling and consumption-based pricing are both increasingly standard, for instance.

The increased prevalence of the cloud has also resulted in something of an arms race between streaming vendors, where there is an increasing tendency for vendors to claim that their product – and, more often than not, their product alone – is "truly" cloud-first or cloud-native. What that actually means tends to vary depending on who you ask, but it is frequently a matter of being designed with the cloud in mind (which is certainly a good thing, but ultimately immaterial unless it manifests itself in a concrete way), being built to take full advantage of the cloud's unique features (such as, say, extremely elastic deployment), and/or being able to provide a highly automated and fully managed service. While all of this can certainly be valuable, the nebulous nature of these differentiators, the multiple and subjective meanings that are used by vendors, and the frequency that vendors seem to claim uniqueness in this regard suggests that you should pause and dig a little deeper whenever you're told that a streaming product is "cloud-native".

There can be negative performance implications for processing or analysing data on the cloud, as opposed to on-prem deployments where network speed is not a factor. However, these environments are in the minority. You may also want, or need, multiple, geographically separated clouds if your streaming data consists wholly or partially of geographically dispersed PII (Personally Identifiable Information). It may be worth considering that some environments –

IoT environments, to be specific – can suffer from poor connectivity (although this is not always the case), which could pose a challenge when it comes to getting data out of the sensor and into the cloud as quickly as possible. This (among other things) advantages vendors that can process and/or analyse data at the edge and move it to a central (cloud) location if – and only if – it is actually useful.

It's also worth noting that the popularity of cloud appears to have largely supplanted yesteryear's fascination with big data and data lakes, which have mostly fallen out of favour in the popular consciousness. At the same time, cloud has grown in popularity less because of this fact, and more because of its advantages in providing flexible, scalable, cost-effective, on-demand, and externally managed software deployments. As alluded to above, this has certainly spurred on the adoption of stream processing. On the other hand, the decline of big data has had – perhaps surprisingly – little impact on the space. Although big data was previously a significant driver for it, in practice the cloud has largely taken its place as the de facto repository for streamed data.

Curation and governance

The need to curate and govern your streaming data, particularly as it enters your system, is seeing increased emphasis within the space. Proactive data governance helps you to maintain high levels of data quality in your streaming data by allowing you to curate it immediately after – or even immediately before – you ingest and store it, which can be extremely important when handling massive quantities of data: polluting your system with poor quality and often opaque data can easily lead to the equivalent of a "data swamp" scenario, where you have a lot of data but you have no idea what any of it is or what it means. Indeed, visibility into your data is a significant component of data governance, particularly via features like data lineage and data catalogues, and this is just as true for streaming data as it is for any other kind.

It's also worth noting that, like the vast majority of data, streaming data needs to be governed in order to comply with recent data privacy legislation (GDPR et al.) and to prevent you from leaking sensitive customer information and breaching consumer trust. We have not seen this mentioned by many vendors in the space, but that doesn't make it any less important. We exhort you to be aware of this when choosing your streaming solution.

Open-source technology

As already discussed, open-source technology has a long history within the streaming space, and has been a driving force in its increasing popularity and adoption. Projects like Apache Kafka, Apache Flink, Apache Spark and Apache Pulsar remain popular and continue to influence the space around them, both in their direct adoption and in their incorporation into proprietary solutions. Kafka in particular remains prominent, to the point of spurring on technologies like Kafka-on-Pulsar (which largely does what it says on the tin).

That said, what once seemed to be a flood of new open-source streaming projects has slowed to a trickle, with few new open-source streaming efforts manifesting over the past couple of years. There have also been several recent acquisitions of major vendors that supported these projects, which we discuss below.

There are two factors to consider here. One is that organisations may have woken up to the difficulties of open source, or more specifically the difficulties (and complexities) of assembling your own streaming solution out of several open-source products. The idea that open source does not always mean low TCO may have finally taken hold, and this has likely combined with streaming's increased popularity to generate a greater willingness to spend money on it.

The other is that data lakes – themselves largely driven by open-source tech – have fallen out of favour, giving way to cloud environments on AWS, GCP and Azure that ultimately serve the same purpose, albeit with different technology and nomenclature. Since these clouds all offer native streaming solutions of their own, there is little perceived need to invest in a separate streaming solution unless you find those solutions inadequate for your needs. In which case, the obvious next step is a suitable proprietary solution, not open source.

Essentially, we posit that cloud solutions have taken the place (or, perhaps more accurately, will soon take the place) of open source as a way for organisations to take their initial steps into the world of streaming without needing to commit large sums of money up front. Open-source technology itself is still alive and kicking – there are several proprietary streaming efforts leveraging it, for instance, let alone home-grown solutions that have already been established – but the period where the greatest competition for any streaming vendor was invariably a DIY Kafka stack is over (and if not, it soon will be).

Vendors

Before describing the vendors we have chosen to include in this report, we should note that it is representative, not comprehensive. In other words, we have included the products and vendors that we feel best exemplify the stream processing space and the strengths and possibilities therein, as opposed to attempting to catalogue each and every available solution. We have ignored products based on offerings we are already covering, and we only cover proprietary solutions, since purely open-source projects generally do not work as streaming solutions on their own, and even if they do, they are not easily comparable to commercial products by their very nature.

Commercial solutions built using open-source products, on the other hand, are fair game. For instance, Confluent, Ververica, and StreamNative all act as enterprise support for – and, more often than not, are the greatest contributor to – a particular open-source Apache project: Kafka, Flink, and Pulsar, respectively. Confluent is a veteran of the space, a lead contributor to Apache Kafka, and offers a streaming platform that leverages it. Ververica provides a similar treatment of Apache Flink, and was previously known as data Artisans, but in the past few years was acquired by Alibaba and rebranded. Nevertheless, it continues to operate as it always has. Streamlio, previously the company behind the open-source platform (built on Apache Heron, Pulsar and BookKeeper) of the same name, is now part of Splunk, which seems content to leverage the technology as part of its own platform, rather than as a discrete streaming solution. However, Streamlio's spirit lives on in StreamNative, which is, similarly, an open-source platform built on Pulsar. More to the point, much of Streamlio's talent and expertise have gone over to StreamNative as well, including its co-founders.

EsperTech is similar to the aforementioned vendors in its support for Esper, another open-source solution, though in this case not an Apache project. It is also interesting in that it offers streaming analytics in addition to its stream processing functionality. The same is true of Striim, and for clarity – and to avoid comparing apples and orange trees – we have appropriately colour-coded these vendors on the Bullseye diagram.

Striim, Informatica, and Push Technology also differ from the other vendors featured in that they are not driving forces behind a particular open-source project, at least in the same way that, say, Confluent is for Kafka. That said, this doesn't mean that they don't strive to leverage open source: far from it, in fact. Informatica, for instance, heavily leverages Apache Spark and Apache NiFi, and is quite fervent in promoting open-source technologies in general. Likewise, Push Technology's Diffusion platform includes a gateway adapter for Apache Kafka, allowing it to readily ingest, parse, and transform Kafka data.

We have omitted streaming solutions from the three major cloud vendors (Amazon, Microsoft and Google). They are certainly viable offerings, albeit offerings exclusive to each individual cloud (which, it should be said, is often a significant disadvantage), but ultimately, we expect readers to fall into two camps: either you have no interest in these clouds whatsoever, or you are already on one or more of these clouds but find their solutions inadequate. We find it unlikely that anyone would migrate to the cloud, have interest in a streaming solution, and not reach for the most immediate and accessible solution available to them. Therefore, including these solutions in their entirety would serve little purpose. However, we will summarise our findings here: these offerings are excellent gateways to the world of streaming, but lack some of the sophistication of many of the other products we have included. We urge you to try them out if you have easy access to them, but be ready to move on if they cannot meet your needs.

Conclusion

Stream processing has benefitted enormously from the rising popularity of the cloud: the cloud frequently makes excellent use of streaming capabilities, while at the same time making it easier than ever to adopt streaming technologies. The popularity of streaming has risen significantly in return.

In fact, it is our belief that real-time data – and hence real-time data processing – is increasingly becoming the de facto standard. More and more companies are no longer asking "why should we do this in real-time?", but rather "why shouldn't we?". This puts stream processing in very good stead: it is fast becoming an essential technology. In summation, there has never been a better time to make use of streaming technology and, consequently, stream processing.

About the author

DANIEL HOWARD
Senior Analyst,
Information Management and DevOps

Daniel began his career in the IT industry relatively recently, in only 2014. Following the completion of his Masters in Mathematics at the University of Bath, he started working as a developer and tester at IPL (now part of Civica Group). His work there included all manner of software development and testing, usually in an Agile environment and usually to a high standard. In the summer of 2016, Daniel left IPL to work for Bloor Research as an analyst, and the rest is history.

Daniel works primarily in the data space, though he dabbles in development, testing, and DevOps. The former often (though far from always) involves working alongside his father, Philip Howard, while the latter allows him to leverage the technical expertise, insight and 'on-the-ground' perspective garnered through his old life as a developer to good effect.

Outside of work, Daniel enjoys latin and ballroom dancing, board games, skiing, cooking, and playing the guitar.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.

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**