MarketUpdate

Streaming Analytics

Market Basics

Streaming analytics is a space that is largely built on the back of stream processing. In turn, stream processing solutions – broadly speaking – exist to ingest, move and/or transform streaming data, and hence tend to focus on data integration and data movement. Streaming data, then, is data that is generated (and hence must be processed) continuously from one source or another. Streaming analytics solutions take streaming data and extract actionable insights from it (and possibly from non-streaming data as well), usually as it enters your system. They may or may not offer stream processing functionality as well.

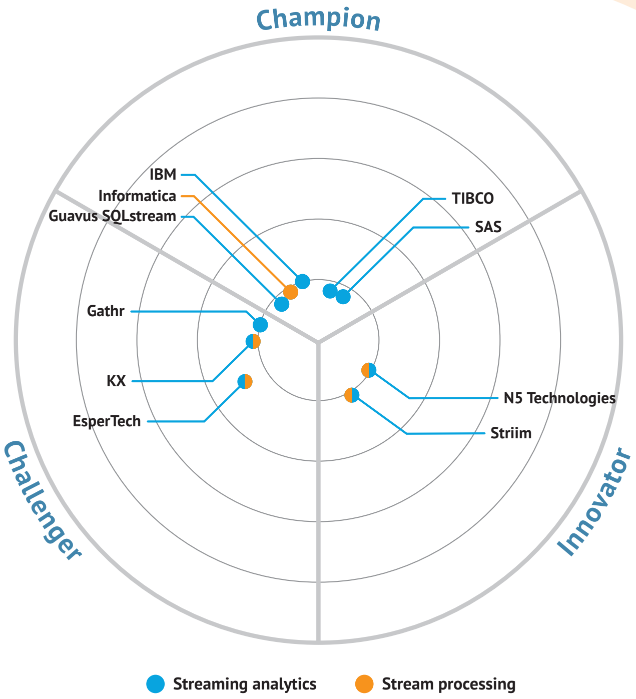
The core idea driving streaming analytics is that you will often benefit from being able to act on streaming data immediately, rather than with a significant time lapse. This is largely because the kind of data that is generated continuously is almost necessarily volatile (otherwise, why generate it continuously in the first place?) and thus particularly benefits from quick action. Streaming analytics enable this by analysing data in real-time as it flows into your system, in turn allowing for faster and more informed responses to that data. This will often be accompanied by dynamic (and again, real-time) visualisations in order make any findings easier to process. Other capabilities, including Business Intelligence (BI), machine learning, data preparation, and so on, are sometimes offered in some augmentative capacity as well.

The space also has a particularly notable open-source presence, to the point that we consider it not just a major trend (as we normally would) but effectively foundational for the space (hence why discuss it in this section as opposed to – or rather, as well as – the next one). Apache projects such as Flink, Pulsar, and most of all 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 analytics 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, as well as machine and deep learning libraries, 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 (Confluent, for example, does so for Kafka), but that instead removes one of the key draws for open source.

Proprietary streaming analytics solutions, by contrast, are frequently designed to offer holistic solutions out of the box. Accordingly, they demand far less effort on your part. Some vendors, such as Impetus, even offer solution stacks built on open-source software, effectively combining several projects together into a complete and comprehensive package. It is not difficult to see why someone might prefer this sort of solution over building their own, even though the advantages of open source can be substantial.

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), and 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 stream processing and analytics. And with exponentially more streaming data 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. 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 (and streaming analytics by extension) has gone from burgeoning – but still essentially niche – to mainstream.

We have identified several discrete trends within the streaming analytics 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 machine learning and AI, of containers, and of IoT is impacting almost every space within data management. Streaming analytics 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 analytics. As you might expect, this is particularly true for the kind of highly performant and scalable streaming solution that can readily handle sensor data arriving at a massive scale.

Machine learning also has history within the space, and the capabilities on offer to support it are particularly intriguing. Support varies wildly between products, with some offering the bare minimum while others make it a core pillar of their solution. Building, hosting and training models (the latter on streaming data specifically) are all capabilities that could be on the cards, as well as some degree of model management and various features for helping you to apply models to your streaming flows.

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 streaming 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 analytics 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.

On a more pessimistic note, there may be negative performance implications for analysing data on the cloud (as opposed to on-prem), although to an extent you can make up for this by simply throwing more money at the cloud in order to get at greater processing power. This approach is not terribly economical, however. 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 also be worth considering that some environments – IoT environments, to be specific – may suffer from poor connectivity, which could pose a challenge when it comes to getting data out of the sensor and into the cloud for analysis as quickly as possible. This (among other things) advantages vendors that can 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. This has had – perhaps surprisingly – little effect on streaming analytics. 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.

Integration with traditional analytics and batch processing

Despite – or perhaps because – of the increasing popularity of streaming analytics, several vendors have accelerated their efforts to integrate their streaming analytics solutions into broader analytics suites, in an attempt to address your general analytical needs more holistically. In some exceptional cases this is through offering all of those analytics themselves, but usually this is more a matter of integrating with third-party solutions. In any case, this approach makes sense: analytics is all about generating actionable insights, and analysing and cross-referencing multiple kinds of data at once can, at least in principle, make those insights more accurate and well-informed.

The most obvious example of this is the now effectively universal incorporation of batch processing into streaming solutions: basically, being able to analyse batch and streaming data simultaneously (or at least through the same interface) is so ubiquitous that it's almost table stakes. An interesting outcropping of this idea that some vendors are offering is the ability to analyse batch data (or streaming data that has already been ingested) as if it were streaming data: in other words, in a time-sensitive manner.

Similarly, in past reports we've discussed the introduction of kappa (as opposed to lambda) architectures as a means to effectively amalgamate streaming and batch pipelines. This discussion is now largely over: kappa architectures have taken hold, and are firmly entrenched in the streaming space. Whereas a few years ago lambda was the standard, and offering a kappa architecture was a notable outlier to the vendor's benefit, it is now the reverse: kappa is the standard, and only offering lambda puts you at risk of falling behind. Essentially, kappa architectures are increasingly an expectation, not a differentiator. At the same time, some vendors are providing compelling alternatives to both lambda and kappa architectures. N5 Technologies' *"Micro DataServices"*, for instance.

Curation and governance

The need to curate your streaming data, particularly as it enters your system, is seeing increased emphasis by a number of vendors (and notably, Confluent has recently announced a dedicated solution suite for governing data in motion). This helps you to maintain high levels of data quality in your streaming data by curating 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. Analysing your data before storing it also has the added benefit of allowing you to throw it out of if you don't want or need it, keeping your data stores clean and reducing storage costs.

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, and Spark Streaming remain popular and continue to influence the space around them, both in their direct adoption and in their incorporation into proprietary solutions. 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 a DIY Kafka stack is over.

Vendors

To start with, we should note that this report is representative, rather than comprehensive: we have chosen to include products that we feel best exemplify the streaming space and the strengths and possibilities therein, obviously emphasising streaming analytics in particular. A couple of more processing-oriented vendors have made it in, due to qualities we feel are exceptionally relevant and therefore worth highlighting. Most notably, Informatica is largely focused on data integration and stream processing, not analytics, but also offers robust AI and machine learning functionality. Others, such as Striim, emphasise data integration and stream processing while also offering substantial analytics capabilities. As such, for clarity – and to avoid comparing oranges and orange trees – we have chosen to colour-code vendors on the Bullseye diagram according to these capabilities.

We have ignored products based on offerings we are already covering (for instance, Oracle Complex Event Processing, which is based on Esper) except when the product is sufficiently distinct from its base (for example, by combining several different technologies together) and we only cover proprietary solutions, since purely open-source projects generally do not work as streaming analytics solutions on their own, and even if they do they are not easily comparable to commercial products by their very nature. Products built using open-source technology are fair game, however.

There have been some significant acquisitions since we last covered this space. Notably, both data Artisans (a vendor that offered commercial support for Flink, now rebranded as Ververica) and Streamlio (an open-source platform built on Apache Heron, Pulsar and BookKeeper) have been acquired: by Alibaba in the case of data Artisans/Ververica and by Splunk in the case of Streamlio. In contrast, Confluent (provider of commercial support for Kafka, and the third major vendor focused on commercialising open source) has not only not been acquired, but was purportedly so busy with sales that they didn't have time to brief with us. Take from that what you will.

In addition to data Artisans rebranding itself Ververica, Impetus has rebranded its 'StreamAnalytix' product to 'Gathr', though the company itself remains Impetus. This is largely to recognise the increased breadth of offering it has been providing in recent releases, and to avoid the implication that it is designed exclusively for streaming analytics. Gathr is also notable for being the exception to one of the ground rules we've laid out: it is, in fact, built around Esper, which we cover separately, but it combines it with Spark Streaming and Apache Storm to become very much its own thing.

That said, we have chosen not to cover any of the three vendors mentioned above in this report. Although they offer competitive streaming solutions, they are all oriented primarily around stream processing, as opposed to streaming analytics. Moreover, they all primarily provide commercial support for freely-available open-source technology. This is a significantly different business model from the other vendors we cover. Finally, our reckoning is that open-source streaming solutions are on a downturn. All of this has combined to make them a poor fit for this report, though, it should be said, not necessarily for streaming as a whole.

We have also 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, 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, but 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 fulfil your needs.

One other notable vendor we have not included is Software AG. Its product, Apama, actually provides a very respectable streaming solution, but Software AG itself appears extremely reticent to market it as such, preferring to focus exclusively on IoT despite the product's substantially greater breadth. We are frankly at a loss as to why the company has chosen to do this (or, more accurately, we understand the reasoning, but don't agree with it).

Finally, a note on IBM Streams, now primarily offered as part of IBM Cloud Pak for Data. Although we continue to include it in the report, this is largely due to a) its legacy as part of the streaming space and b) its widespread usage. It is difficult to argue that IBM isn't a leading vendor in the space purely for these factors. Likewise, its streaming products are competent and well put together. However, as far as we can tell, it has done little – if anything – to develop its capabilities in the three years since our previous report. Even accounting for loss of work due to COVID, this is very disappointing. Not only is this approximately an eternity in software years, the space itself has been anything but static in that time. Our opinion is that, at least when it comes to streaming, IBM is resting on its laurels, and if it were not such a major player, we would drop it from the report entirely.

Metrics

To score the various vendors/products discussed in this report we have used the following metrics, largely inherited from the predecessor to this report:

- **Analytics and modelling** – the extent to which the platform supports analytics, and particularly AI-driven analytics, either as embedded functionality within the product or integration with third party tools and libraries. Issues would include whether models can be trained within the platform or only outside of it, the extent of support for models built in various languages, and the ability (or lack thereof) to apply models to, or integrate models with, streaming data pipelines. Data preparation capabilities built into the platform are also relevant, as is the ability to provide streaming analytics in combination with more traditional analytics functionality, whether by providing that functionality directly or via integration with other solutions.
- **Development** – how easy is it to develop applications and/or analytics using the tools (if any) that are provided? This will include considerations such as whether there is a visual development environment, whether a common IDE such as Eclipse is available, and whether language training (for example, SQL – particularly ANSI SQL – versus a proprietary language) is required.
- **Architecture** – how easy is it to scale the solution? Is the platform capable of handling tens of millions of events per second? Millions? Or hundreds of thousands? Further, what is the footprint of the solution: is it suitable for deploying in edge devices or gateways?
- **Deployment** – what platforms does the product run on? Is it available both in-cloud and on-premises? What administrative tools are available? How easy is the process of deployment? How well, and to what extent, does the product support cloud deployments, and does it offer features that are notably conducive to the cloud, such as dynamic scaling? Also, what facilities are provided to monitor streams flowing through the environment as well as the performance of the cluster underpinning the solution?
- **(Non-analytic) streaming functionality** – going beyond analytics, to what extent does the platform support data integration and transformation functions? Does the product do anything to support the curation or governance of streaming data and processes? Does it work with batch as well as streaming data? Are there workflow capabilities built into the product? Does the product support *"exactly once"* processing? Is it event-based or window-based and, if the former does it also support time windows? Does it support functions such as tumbling windows, sliding windows and so forth?
- **Connectivity** – how extensive are the connectivity options for IoT sources as well as more traditional connectivity requirements? Also within this category would be the range of data types supported: for example, does the product extend beyond structured and semi-structured data? Does it support text, voice and so forth? API support to access machine learning libraries is also relevant.
- **(Breadth of) integration** – to what extent is the platform integrated with other solutions, either from the same or third-party vendors. In other words, is this part of a larger solution stack with significant complementary capabilities? If so, how comprehensive is that? Does the product include integration with third party (or provided) databases for storing event and other forms of data, and how good are the facilities for combining event analytics with historic data stored in a data warehouse, mart or lake? To what extent has the product been designed to play a role in a larger stack?
- **Self-service** – how amenable is the platform to use by business analysts? Are there self-service and collaborative capabilities built-in? Are there visualisation capabilities provided and/or is there connectivity at the front-end to support visualisation tools such as Tableau?

Positioning on the Bullseye diagram also encompasses factors agnostic to the streaming space (innovation, say), as well as company issues such as support, geographic presence, stability and so on.

Conclusion

The current climate in information management demands that every product and every space must be judged, to one extent or another, by its contribution to the three most significant trends of the day: the cloud, AI and machine learning, and IoT. Streaming analytics has hit its stride and taken its place in the mainstream in part because it addresses all three: IoT has been a significant driver (arguably the significant driver) of the space for years, machine learning has always been a core component of it, and the cloud can make excellent use of streaming capabilities while making adoption of streaming analytics easier than ever. In summation, there has never been a better time to make use of streaming analytics.

About the author

DANIEL HOWARD
Senior Analyst,
Information Management and DevOps

Daniel started 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 and web development and testing, usually in an Agile environment and usually to a high standard, including a stint working at an 'innovation lab' at Nationwide.

In the summer of 2016, Daniel's father, Philip Howard, approached him with a piece of work that he thought would be enriched by the development and testing experience that Daniel could bring to the table. Shortly afterward, Daniel left IPL to work for Bloor Research as a researcher and the rest (so far, at least) is history.

Daniel primarily (although by no means exclusively) works alongside his father, providing technical expertise, insight and the 'on-the-ground' perspective of a (former) developer, in the form of both verbal explanation and written articles. His area of research is principally DevOps, where his previous experience can be put to the most use, but he is increasingly branching into related areas.

Outside of work, Daniel enjoys latin and ballroom dancing, 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**