Streaming Analytics

MarketUpdate

Introduction

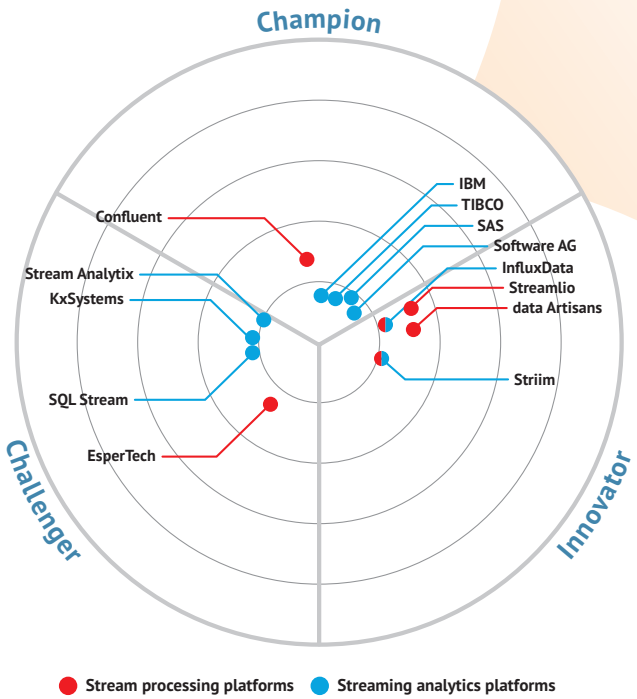
This Market Update both extends and references our 2016 Market Report on Streaming Analytics. It references it to the extent that we do not here go into the specific features that you might expect from a streaming analytics platform, as they were covered in that paper (see ***https://www.bloorresearch.com/technology/streaming-analytics-platforms/*** for more details); and it extends our previous report because we have broadened our coverage to include stream processing platforms.

In that context, it is important to understand how stream processing differs from streaming analytics. The first and probably most significant difference is that many stream processing products are targeted at data integration and data movement as well as or instead of analytics. While we will discuss this element of stream processing our primary focus is on support for analytics. The other major difference is that streaming analytics platforms are typically proprietary and usually encompass a much broader product set than stream processing products which are, generally speaking, more narrowly focused, and which consist of open source, mostly Apache projects. For example, as one vendor in the latter space explained to us *"we offer a streaming processing platform that supports analytics but we do not offer a streaming analytics platform per se."* Thus streaming analytics platforms are often delivered with tightly integrated or embedded business intelligence and analytics tools, may have data preparation capabilities built into them, and you can often train machine learning models on the platform. They are, in effect, solutions rather than platforms for solutions.

In effect, streaming analytics vendors tend to offer solutions whereas the vendors supporting stream processing products are offering one or more components of a solution. More particularly, vendors in the former camp tend to focus exclusively on analytics whereas stream processing suppliers target both data integration and analytics. In our Bullseye diagram we have therefore colour coded these separately so that readers can compare apples with apples. As usual, there are exceptions and these are bi-coloured in our diagram. However, this has still left us in something of a quandary because there is a third class of supplier, which have put together solution stacks based upon the various open source stream processing projects.

Most notably, this includes the three Hadoop distributors. We have elected not to include these vendors, primarily because their stream processing capabilities are discussed here anyway and it would be complex to compare Confluent (which provides commercial support for Kafka) with, say, Hortonworks, which also leverages Kafka. The exceptions to this rule are where vendors are unique in supporting a particular technology or where open source products have been leveraged as a part of a proprietary solution.

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.



Key: products are colour coded to ensure that readers do not compare apples with pears. Note that all vendors in this Bullseye are consider best-of-breed.

The stream processing ecosystem

As noted, stream processing platforms exist within an ecosystem – typically based on Apache projects – that can be used together to build a streaming analytics solution. The fundamental elements of this ecosystem comprise data flow management, distributed messaging, stream processing and machine and deep learning libraries such as MLlib, MADlib and TensorFlow. Associated technologies include Apache Beam and assorted databases such as Apache Druid and Apache Bookkeeper.

In terms of data flow management this is essentially similar to the sort of data flow management you might find in a data integration tool. For example, Apache NiFi is typically used to manage data flows within an Internet of Things (IoT) environment, together with Apache MiNiFi, which focuses on the collection of data at source, particularly on edge devices (using an agent-based approach). Competitive to this is Streamsets, which also has overlapping capabilities with distributed messaging.

The most well-known Apache-based distributed messaging system is Kafka. However, while this was the space it originally started in, the technology has since has been expanded to incorporate processing and storage capabilities. Also in this category are Akka (which is open source but not an Apache project), Apache Flume and Apache Pulsar (which is an incubator project).

As far as streaming platforms themselves are concerned, there are multiple choices. Offerings include the Apache projects: Apex, Flink, Heron (incubating), Kafka, Samza, Spark Streaming and Storm as well as the non-Apache projects Akka Streams and Gearpump (based on Akka).

Apache Beam is an abstraction layer for stream processing that lets you write code against a standard API and then execute the result on any platform.

Finally, there are various types of complementary database that may be associated with stream processing. Most commonly, these are either event stores (for log and similar data) or OLAP (online analytic processing) databases implemented via SQL on Hadoop engines.

The vendors and products

This is intended to be a representative report rather than a comprehensive one. In so far as open source products are concerned our reviews are limited to the discussion of the vendors that are the main contributors to, and which provide commercial support for, the various Apache projects we cover. Thus the vendors concerned are Confluent (Kafka), data Artisans (Flink), and Streamlio (Pulsar, Heron and Bookkeeper). We would have included Databricks had they responded to our requests for information but we do provide a brief discussion of relevant Spark features and, in any case, StreamAnalytix extensively leverages Spark Streaming (as well as Storm). We have excluded Samza and Storm because our research makes it clear that these products have lost traction within the marketplace. We have also omitted Gearpump as this remains only an incubator project. The one company that we might have included but have not, and which has perhaps drawn the short straw, is Lightbend, the company behind Akka and Akka Streams.

As far as proprietary vendors are concerned, we have taken a similar approach as to open source projects in the sense that we have omitted products that are based on other offerings that we are covering already. Thus neither Oracle Complex Event Processing nor Amazon Kinesis are discussed, because they are both based on third-party technology (Esper and SQLStream respectively, where Esper is an open source – but not an Apache project – complex event processing engine). We have, however, included StreamAnalytix, which is built on Esper, because it also leverages both Spark Streaming and Storm, as discussed above.

In our 2016 report we also included Cisco, Microsoft, Time Compression Strategies (Ancelus) and Informatica. In the case of Informatica it is true that the company has an event-process engine but this is used to support data integration rather than analytics and so has been omitted from this paper. The same consideration would apply to other data integration vendors, such as Ab Initio, that have event processing capabilities.

Cisco, Microsoft and Time Compression have all drawn short straws because of the number of new vendors that we wanted to include here. This should not be treated as any aspersion on their respective products and the same is true for Oracle. It is likely that their relative positions compared to other streaming analytics platforms has remained relatively constant over the last two years and readers interested in their positioning should refer to our previous Market Report.

As far as Ancelus is concerned,we have replaced this with InfluxDB. The key issue with database-oriented approaches to streaming analytics is that they support time-series. Ancelus, InfluxDB and Kx Systems all do this. However, if you look at www.db-engines.com (not necessarily the fount of all wisdom) you will see that InfluxData, the company behind InfluxDB, is generating the most interest as a time-series database and we have therefore opted to include InfluxDB in place of Ancelus. Again, this should not be considered as any aspersion on Ancelus.

Market trends

There are multiple trends within this marketplace and rather than place them all within a single section we have broken these down into three segments. We will, however, comment that we have not seen much (any) consolidation in this space. We do not believe this can continue and we expect to see acquisitions in both the open source and proprietary sections of the market.

Apache projects

Evolution in the Apache space is noticeable: projects that used to get a lot of press – for example, Storm – no longer do so. On the other hand, new projects continue to arise and their supporters will claim that these represent next generation capabilities that go beyond earlier products. Thus, Streamlio will claim that Pulsar is a step forward from Kafka and that Heron goes beyond Spark Streaming, Flink or Apex. It remains to be seen whether the market accepts this proposition but the trend is clear that only very few of the Apache projects can sit on their laurels because challenge is constant.

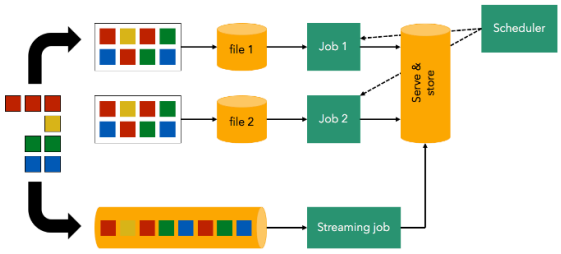
The preceding comments have been justified only just recently. We had originally intended to include DataTorrent, the company behind Apache Apex, in this report. However, in May 2018 it closed it's doors, despite having raised over $20m in funding. This probably means that Apex is a dead duck. And it should be a warning about the potential frailties of companies in the Apache space: we do not expect that DataTorrent will be the last of these to go out of business.

More specifically, there is a clear trend towards vendors extending their offerings to incorporate batch as well as stream processing and to target the data integration space. While we do not think that these products are in a position to replace traditional data integration vendors such as Informatica and Ab Initio (both of which have event processing capabilities of their own) they may be able to supplant them for particular use cases. More generally, however, they are (successfully) targeting new deployments rather than replacements. We should also add that Striim (a proprietary vendor) is also targeting the data integration market, as well as analytics.

Lambda and Kappa architectures

Going beyond this is the debate – largely over – between Lambda and Kappa architectures. A Lambda architecture, illustrated in ***Figure 1***, consists of a streaming pipeline, one or more batch pipelines, and a (database) server to unite the data from these two environments. A Kappa architecture conflates the streaming and batch pipelines into a single environment, thereby significantly simplifying any deployment. All our research suggests that Lambda is being replaced by Kappa though, sadly, some vendors seem not to have woken up to this fact.

In conjunction with the Kappa architecture, vendors are increasingly deploying databases as a part of this architecture. Thus Kafka now includes serialised storage and Streamlio offers Bookkeeper in conjunction with Pulsar and Heron. Other vendors in the open source community are relying on Apache projects such as Druid, while proprietary vendors



are adopting a similar approach with, for example, StreamAnalytix integrating with Kyvos (both products are developed by Impetus Technologies), IBM introducing its Event Store and Software AG leveraging Terracotta DB.

Spark Streaming

Spark Streaming has been historically been *"window"* based. That is, based on a time window. In other words, processing data as a micro-batch. The problem with this is that **a)** you cannot process individual events, **b)** you have issues with events arriving out of time sequence and **c)** you have problems with long running events where you have a start and end time that need to be correlated with one another. As a result, Spark Streaming is not suitable for many streaming applications. The Spark community has started to rectify this though true event based processing will not be introduced – that is the plan – until version 2.4 (version 2.3 was released in March 2018). In the meantime, *"structured streaming"* is available and this provides facilities for handling things such as late arriving data. This is a start but Spark has still got some way to go.

As we have noted, we would like to have been able to include an evaluation of Databricks in this report. Failing that, we have considered including an evaluation of Spark Streaming based on publicly available information. This would, however, be unfair on Spark as the commercial supporters of Flink and Heron have added additional capabilities that go beyond the stream processing platform itself. For what it is worth we would comment that while Spark has both greater maturity and a much larger installed base than any of its Apache-based competitors, we believe that these vendors have a technological edge when compared to Spark Streaming.

Streaming analytics platforms

Perhaps the most notable trend in this space is based on the acquisitions being made by Software AG and TIBCO. Both of these companies – with two of the most mature products in the market with Apama and StreamBase respectively – have been rounding out their product portfolios by acquiring companies with complementary technologies. For example, Software AG acquiring Cumulocity, which proves connectivity to many sensors and actuators; and TIBCO acquiring Statistica and Alpine Data Labs.

Database-oriented streaming analytics

An important point to make is that it tends to be assumed that streaming analytics involves processing millions or tens of millions of events per second. This is not necessarily the case. As an example, the European Space Agency uses InterSystems to process satellite telemetry data but this is at around 100,000 transactions per second rather than much faster rates. The truth is that there are many Internet of Things (IoT) environments where this sort of scale is perfectly adequate and where suitably scalable and performant database offerings may be appropriate deployment options. Readers with less onerous processing requirements would be well advised to consider the deployment of database-oriented solutions.

In particular, there are more and more database vendors targeting streaming analytics and stream processing environments, especially in HTAP (hybrid transactional/analytic processing) environments where you want to combine analytic processing on real-time and historic data. This market is being targeted by both mainstream vendors (SAP, Oracle, IBM and Microsoft) but also various SQL on Hadoop engines (for example Esgyn) as well as others such as InterSystems IRIS, Splice Machine and VoltDB. In the context of this report it is also worth referencing Terracotta DB, which is an HTAP engine from Software AG, and which integrates with the company's Apama streaming analytics platform.

Increased support for time series data is also relevant: db-engines (*www.db-engines.com*) lists 23 time-series databases but does not mention Ancelus, IBM Informix, InterSystems or OneTickData, for example, all of which support time series and which might or would be suitable for supporting streaming analytics. We have included both InfluxData and Kx Systems as exemplars of time series databases in this report because we believe that they can genuinely compete – in performance terms – with other streaming analytics products detailed in this Market Update.

In the context of time series databases it is worth commenting that InfluxData not only targets IoT environments but also IT operations and infrastructure. For example, by monitoring application requests in a microservices environment you can use InfluxData to automatically scale up the number of containers in use. Not only is this difficult to do using Kubernetes it is a use case not being addressed by other vendors in this report.

Algorithmic models and machine learning

There is a wide disparity in the support for algorithmic modelling and machine learning provided by the different vendors. One provider only supports Java models, others support Python, Scala or R but not Java. It is therefore difficult to identify trends though Python and R seem to be most popular.

There is an interesting divergence between providers of proprietary streaming analytics platforms and suppliers of open source stream processing. We are seeing substantial take-up of PMML (predictive modelling mark-up language) to enable the portability of models amongst the former but not from the latter. This may be to do with the fact that streaming analytics products can often train models on their platform, while stream processing offerings cannot. This is not universal: it is typically the case that only some models can be trained as well as scored in-platform but we expect the number of model types that can be so trained to grow. We also anticipate that the range of models supported by PMML will increase. As a result, it seems likely that the stream processing providers will need to extend their platforms to support in-platform training and scoring in order to compete with proprietary solutions.

Scoring

To score the various vendors/products discussed in this report we have used the following metrics:

- **Analytics and modelling –** measures the extent to which the platform supports 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 Java, Python, Scala, R and so forth, as well as import via PMML. Data preparation capabilities built into the platform are also relevant.
- **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 versus a proprietary language, is required). Data preparation capabilities built into the platform are also relevant, as is workflow.
- **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? 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 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 are 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?

We recognise that some aspects of these requirements will be more important for some users than others. So, while all of the scores for individual products are included in the detailed descriptions that follow later, the tables below represent the comparative scoring for each of the areas set out above. Note that each score is out of 5 but, unlike Amazon or Trip Advisor, it is impossible to score 5 on any topic. A score of 5 would represent a *"perfect"* product at this time. As we do not believe in perfection no product can be awarded a maximum score.

We have split the scores across stream processing platforms and streaming analytics solutions. However, Striim and InfluxData appear in both sections because, while they both offer solution-oriented products they also target markets (data integration and IT operations respectively) that are not specifically about streaming analytics.. We have also included EsperTech in the former category because its Esper and NEsper (the .Net version of Esper) products, although complex event processing (CEP) engines, are intended for deployment within a larger stack and thus are platform rather than solution oriented.

The scores below are solely related to the products under evaluation. However, the positioning on the Bullseye diagram, as well as the *"mutable"* diagrams accompanying each vendor evaluation, also encompasses company issues such as support, geographic presence, stability and so on; as well as factors like innovation and the ability to support moves towards a data-driven enterprise.

Scores for stream processing platforms

VENDOR	ANALYTICS AND MODELLING
EsperTech	3.5
InfluxData	3.5
Streamlio	3.0
Striim	3.0
data Artisans	2.5
Confluent	2.5

VENDOR	NON-ANALYTIC STREAMING
Striim	4.5
Confluent	4.5
data Artisans	4.0
InfluxData	3.5
Streamlio	3.5
EsperTech	2.5

VENDOR	DEVELOPMENT
InfluxData	4.5
Striim	4.0
data Artisans	3.5
Confluent	3.0
Streamlio	3.0
EsperTech	2.5

VENDOR	CONNECTIVITY
Confluent	4.5
Striim	4.5
data Artisans	4.0
EsperTech	4.0
Streamlio	3.5
InfluxData	3.5

VENDOR	ARCHITECTURE
Striim	4.5
data Artisans	4.0
EsperTech	4.0
InfluxData	4.0
Streamlio	4.0
Confluent	3.5

VENDOR	INTEGRATION BREADTH
Confluent	4.5
data Artisans	3.5
InfluxData	3.5
EsperTech	3.5
Streamlio	3.5
Striim	3.0

VENDOR	DEPLOYMENT
Confluent	4.0
data Artisans	4.0
EsperTech	4.0
Streamlio	4.0
InfluxData	3.5
Striim	3.5

VENDOR	SELF-SERVICE
InfluxData	4.0
Streamlio	3.5
Striim	3.5
Confluent	3.0
data Artisans	3.0
EsperTech	2.5

Scores for streaming analytics platforms

VENDOR	ANALYTICS AND MODELLING
IBM	4.5
SAS	4.5
Software AG	4.5
StreamAnalytix	4.0
TIBCO	4.0
Kx Systems	4.0
SQLStream	4.0
InfluxData	3.5
Striim	3.0

VENDOR	NON-ANALYTIC STREAMING
Striim	4.5
SQLStream	4.0
StreamAnalytix	4.0
SAS	4.0
InfluxData	3.5
TIBCO	3.5
IBM	3.5
Software AG	3.5
Kx Systems	3.0

VENDOR	DEVELOPMENT
SQLStream	4.5
TIBCO	4.5
InfluxData	4.0
Software AG	4.0
SAS	4.0
StreamAnalytix	4.0
IBM	3.5
Kx Systems	3.5
Striim	3.5

VENDOR	CONNECTIVITY
SAS	4.5
Software AG	4.5
Striim	4.0
IBM	4.0
TIBCO	4.0
Kx Systems	3.5
SQLStream	3.5
InfluxData	3.5
StreamAnalytix	3.5

VENDOR	ARCHITECTURE
SAS	4.5
Software AG	4.5
TIBCO	4.5
IBM	4.0
Striim	4.0
InfluxData	4.0
Kx Systems	4.0
SQLStream	4.0
StreamAnalytix	4.0

VENDOR	INTEGRATION BREADTH
IBM	4.5
Software AG	4.5
TIBCO	4.5
SAS	4.0
StreamAnalytix	4.0
InfluxData	3.5
Kx Systems	3.5
SQLStream	3.0
Striim	3.0

VENDOR	DEPLOYMENT
SQLStream	4.5
StreamAnalytix	4.5
Kx Systems	4.0
SAS	4.0
Software AG	4.0
Striim	4.0
InfluxData	3.5
TIBCO	3.5
IBM	3.5

VENDOR	SELF-SERVICE
Software AG	4.5
SAS	4.0
SQLStream	4.0
TIBCO	4.0
InfluxData	3.5
Kx Systems	3.5
IBM	3.5
StreamAnalytix	3.5
Striim	3.5

Conclusion

There are no weak products in this market. Many of the offerings have been available for the better part of two decades and they are as mature as you might expect. This is true for most, though not all, of the proprietary vendors, plus EsperTech. It might seem that this was not true of the suppliers offering Apache-based products. Some of these – notably Streamlio – are very new indeed. However, it tends to be the case, even with Streamlio, that its technology has been tested in demanding environments by its original developers. We therefore stand by our statement that there are no weak products in this arena: the principle choice will be between a solution stack on the one hand and a platform upon which you can build a solution on the other.

Bloor

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

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**

About the authors

DANIEL HOWARD
Senior Researcher

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.

PHILIP HOWARD
Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

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

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

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
Email: **info@Bloor.eu**