

Transform your business insights with streaming analytics

Acquire, analyze and act on all your data in real time



1

Streaming analytics 101

2

The state of streaming analytics

3

Evaluating streaming analytics

4

IBM Streams in the real world

5

What it means to you

Throughout history, insights have keyed success. Empires have been created, and businesses have collapsed based on the speed and quality of their insights.

While the power of insights still holds true, today's challenge comes from the massive data quantity that must be mined to obtain those insights. In fact, Forbes predicts by 2020 about 1.7 MB of new information will be created every second of every day—for each and every human being on the planet.¹

Because of the continual flood of new data points and connections, the insights generated from this data can degrade rapidly. That means to maximize their value, you must uncover these accurate insights in real time so you can act on them right away.

This is the power of streaming analytics. Unlike traditional analytics used for after-the-fact analysis, streaming analytics solutions can help you detect insights in high-velocity data streams. That ability lets you detect and assess urgent opportunities, automate immediate actions and dynamically adapt your actions—allowing you to exploit opportunities in real time.



Streaming analytics platforms deliver a level of performance that cannot be achieved at any cost by traditional approaches to data management.

1 Streaming analytics 101

2 The state of streaming analytics

3 Evaluating streaming analytics

4 IBM Streams in the real world

5 What it means to you

Streaming analytics 101

An effective streaming analytics solution requires two key components to push real-time contextual intelligence to applications. First, it must include development tools to create streaming flows that applications can use. Second, the solution must include a scalable runtime platform to leverage those flows.

At the center of the streaming flows rests a sequence of streaming operators. These operators are configured and stitched together to process and analyze the incoming data streams. Understanding these fundamental building blocks is crucial to maximizing the value of your streaming analytics solution.

Transformation

Streaming data delivers tremendous volumes of information, and often this can be filled with irrelevant noise rather than signals. Transformation operators allow developers to narrow the incoming stream to include only data relevant to the application.

Such operators are often the first set applied to streaming data, serving to remove the useless noise from the streaming data. Then filtering operators may filter according to terms, data sources or location. For example, a banking application may want to drop all events generated in the US if it's trying to detect fraudulent transactions in Canadian branches.

“The future of streaming analytics lies at the heart of the real-time enterprise for driving business process automation.”

Ronnie Biggs,
Bloor Market Report “Streaming Analytics 2016”

Correlation

Streaming applications typically combine data from multiple sources. Aggregation or correlation operators allow developers to combine multiple streams into one stream like tributaries flowing into a larger river. For example, a telecom application may need to link incoming call data records with customer profiles to offer an upgrade to a higher-tier calling plan.

Enrichment

Streaming data often requires reference data to provide additional context. Enrichment operators allow developers to pull in reference data from various databases.

This reference data adds context for other streaming operators. For example, a data stream from a home improvement store's point-of-sale (POS) system could enrich its transactions against customer records, revealing that a customer is probably working on a roofing project.

Time windows

Streaming data flows in real time, but applications often need a snapshot of the stream over a discrete time period. Time window operators allow developers to define a time period along with the streaming data to include in that particular time window.

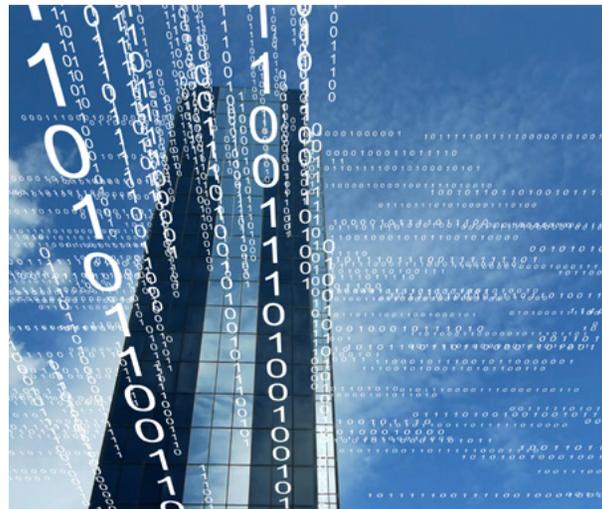
Developers can then use those operators to perform time series analyses in real time, such as running totals, weighted moving averages, Bollinger Bands and others. For example, a time window could show all transactions in the last five minutes that exceeded USD 10,000 in value and calculate a rolling average.

Pattern matching

Streaming data often contains interesting patterns that only emerge as new streaming data arrives. A common pattern occurs when an event A arrives at time t , and another event B arrives at time t plus x .

Pattern operators allow developers to define arbitrarily complex relationships between streaming events. However, it's important to remember there can be a tradeoff between the complexity of the pattern and the latency introduced by needing to process the analytics.

A pattern detection example is traffic analysis. For example, streaming analytics could use temporal patterns to note that a set of vehicles all decelerated on a highway within 10 seconds, indicating a possible accident or the presence of a state trooper.



Business logic

Streaming analytics are designed to supply applications with real-time context. Business logic operators allow developers to define if-then clauses to push contextual information to applications, or provide services that allow applications to query a relevant real-time state.

For example, a peer-to-peer e-commerce site could use streaming analytics to identify illegal behavior among users. Once identified, then the application could immediately send an email or desktop alert to compliance personnel.

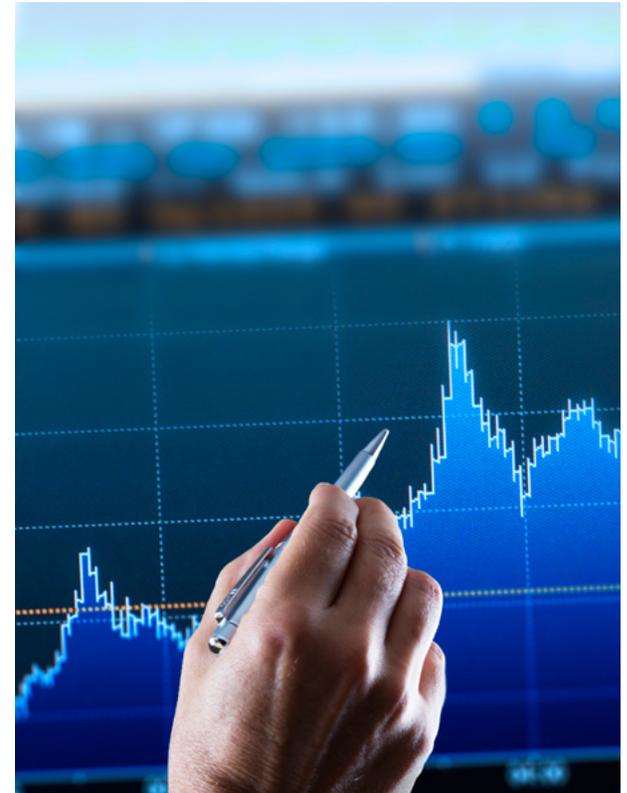
The state of streaming analytics

Because of the wide range of applications, businesses are adopting a variety of approaches to streaming analytics. These approaches include fast in-memory databases with streaming data extensions, as well as in-memory data grids for event-driven streaming data analysis.

This wide variety of approaches is driving other changes as well. Once a developer tool, business analysts have joined the streaming analytics revolution due to graphical tools that allow rapid development without coding.

Further, the Bloor Market Report *Streaming analytics 2016* points out that streaming analytics growth is exploding outside traditional IT niches. It notes the industrial Internet of Things (IoT) and machine-to-machine interaction are accelerating adoption of streaming analytics across multiple sectors.

And there's a clear trend towards integrating streaming analytics with predictive analytics, machine learning and rules engines. The reason is simple – streaming analytics help improve accuracy and reliability, and simplify the expression of complex business patterns and scenarios.



For IoT applications, *Streaming analytics 2016* noted there are many areas where streaming analytics are generating significant business value. When also considering growth potential and competitive pressures going forward, it highlighted the following use cases:

- **Preventative maintenance**—real-time analytics allow reduced operational and equipment cost by minimizing unplanned outages and avoidable site and maintenance visits.
- **Retail**—real-time inventory updates help drive business processes for inventory and pricing optimization, as well as optimization of the supply chain, logistics and just-in-time delivery.
- **Smart transportation**—real-time tracking and vehicle telematics provide operational advantages for companies like Uber, usage-based pricing for tolls and car sharing, location-based insurance applications and driver advisory systems with driver feedback to improve vehicle-operating costs.
- **Smart energy**—real-time monitoring of smart meters permits smart pricing models for electricity, as well as integration with renewable energy generators to optimize power generation and distribution.
- **Industrial automation**—streaming and predictive analytics enable manufacturers to optimize production processes and product quality, including automated alerts and production shutdown when quality levels are breached.
- **Healthcare**—real-time data facilitates integrating a variety of smart sensors to monitor patient condition, medication levels and even recovery speed to optimize care recommendations.

Evaluating streaming analytics

As streaming analytics become more valuable for discovering actionable insights, analysts have conducted comprehensive market evaluations to guide streaming analytics customers. And their approaches vary depending upon their chosen emphasis areas.

The Forrester Wave

In *The Forrester Wave™: Big Data Streaming Analytics, Q1 2016* report, Forrester evaluated the top 15 commercial and open source streaming analytics vendors against 26 different evaluation criteria. These criteria were set based on past research, user requirements and vendor interviews. For ease, it grouped the 26 criteria into the following three categories:

- 1. Current offering.** For this category, Forrester evaluated each vendor's product for workload scalability, ingestion throughput, analytical throughput, fault tolerance, stream handling, streaming operations and application development features. The purpose was to establish the differentiated capabilities of the vendor's current offering as of 31 December 2015.
- 2. Strategy.** For this category, Forrester compared each vendor's overall stream analytics strategy and product road map to assess that vendor's ability to compete and grow in the enterprise streaming analytics market. Key factors included Forrester's confidence in the vendor's ability to execute its stated strategy, as well as support current and future customers.

- 3. Market presence.** For this category, Forrester assessed each vendor's overall market presence. This assessment included evaluating overall streaming analytics revenue, installed base, market awareness of the vendor's product, and partnerships with other technology and services firms.

Each of the 15 vendor's products had a baseline of capabilities in order to be considered. The capabilities included core streaming analytics functionality to provide analytics features on streaming data in time windows. They also included general-purpose streaming analytics products that weren't embedded, and didn't include functionality focused within domain-specific applications.



Additionally, the software solutions had to be capable of installation on a client’s on-premises, private cloud or public cloud infrastructure. Cloud-only solutions were excluded. Finally, each vendor was required to furnish Forrester with at least two customer references who were willing to share their experiences with the product.

When the dust settled, Forrester recognized IBM as an overall leader in big data streaming analytics. IBM’s status was based on its top ranking for the *Current offering* category and its tie for highest score in the *Strategy* category.

“IBM’s architecture can flex to handle any streaming challenge, and the development environment provided one of the richest operator sets in the market.”

The Forrester Wave™: Big Data Streaming Analytics, Q1 2016

Figure 1. The Forrester Wave data streaming analytics, Q1 2016

- | | | | | |
|----------------------------------|---|---|--|-------------------------------|
| 1 Streaming analytics 101 | 2 The state of streaming analytics | 3 Evaluating streaming analytics | 4 IBM Streams in the real world | 5 What it means to you |
|----------------------------------|---|---|--|-------------------------------|

As proof of IBM’s leadership, Forrester cited the fact that IBM® Streams enabled cognitive solutions. This ability allows IBM Streams to incorporate all intelligence, including natural interfaces, situation awareness, smart decisions and learning, to better uncover insights.

Forrester also noted that IBM Streams could ingest and understand the always-on data streams from applications and IoT devices—the key to making cognitive-based decisions. Finally it stated, “IBM’s architecture can flex to handle any streaming challenge, and the development environment provided one of the richest operator sets in the market.”

	Forrester's Weighting	Cisco Systems	data Artisans	DataTorrent	EsperTech	IBM	Impetius Technologies	Informatica	Oracle	SAP	SAS	Software AG	SQLstream	Strim	TIBCO Software	WSO2
CURRENT OFFERING	50%	2.91	3.34	3.65	3.14	4.88	3.20	3.10	3.57	4.37	3.48	4.73	3.60	3.19	3.86	2.79
Architecture	40%	3.00	4.50	4.75	3.00	5.00	3.50	2.25	3.25	4.25	3.50	4.75	3.50	3.50	3.50	3.25
Operational management	5%	0.00	3.00	5.00	3.00	5.00	5.00	5.00	3.00	5.00	1.00	5.00	5.00	3.00	5.00	1.00
Stream handling	15%	3.00	3.00	3.00	3.00	4.20	1.80	3.40	3.40	4.20	3.40	5.00	3.40	3.00	4.20	2.20
Streaming operators	20%	3.60	2.80	2.40	3.80	5.00	3.60	3.80	4.00	4.40	4.20	4.60	3.80	3.60	3.40	3.60
Application development	15%	2.60	2.50	3.80	3.50	5.00	3.70	3.50	4.40	4.40	3.50	4.40	3.50	2.80	5.00	2.60
Business applications	5%	3.00	0.00	0.00	1.00	5.00	0.00	3.00	3.00	5.00	3.00	5.00	3.00	1.00	3.00	0.00
STRATEGY	50%	3.35	2.90	3.95	2.85	4.60	3.80	3.50	4.05	4.60	3.40	4.30	3.95	3.55	4.30	3.65
Acquisition and pricing	20%	1.00	5.00	5.00	5.00	3.00	3.00	3.00	3.00	3.00	1.00	3.00	5.00	3.00	3.00	5.00
Solution road map	30%	3.00	3.00	4.00	2.00	5.00	4.00	3.00	4.00	5.00	4.00	4.00	4.00	4.00	4.00	3.00
Ability to execute	25%	5.00	3.00	4.00	3.00	5.00	4.00	4.00	5.00	5.00	5.00	5.00	4.00	4.00	5.00	4.00
Implementation support	25%	4.00	1.00	3.00	2.00	5.00	4.00	4.00	4.00	5.00	3.00	5.00	3.00	3.00	5.00	3.00
MARKET PRESENCE	0%	0.48	0.25	1.35	2.18	3.80	0.23	1.88	3.05	3.58	1.00	4.10	1.23	1.35	3.95	0.63
Vendor sizing	75%	0.30	0.00	0.80	1.90	3.40	0.30	1.50	2.40	3.10	1.00	3.80	1.30	0.80	3.60	0.50
ISV and services partners	25%	1.00	1.00	3.00	3.00	5.00	0.00	3.00	5.00	5.00	1.00	5.00	1.00	3.00	5.00	1.00

All scores are based on a scale of 0 (weak) to 5 (strong)

Figure 2. The Forrester Wave streaming analytics rankings, Q1 2016

Bloor Market Report

In *Streaming analytics 2016*, Bloor Research International takes a different approach to evaluating streaming analytics offerings. Bloor notes that streaming analytics are an essential component of edge intelligence, and that vendors are now looking at embedded and appliance-based deployments as a complement to their cloud solutions.

For this reason, Bloor’s evaluation criteria differed from Forrester’s because it considered edge processing for the IoT. In fact, this is one of the five criteria the Bloor Market Report considered:

- 1. Streaming analytics features.** For this criteria, Bloor evaluated the richness of higher-order data processing operators, including time windows, transformation, data aggregation and pattern matching, as well as geospatial and location-based capability.
- 2. Solution footprint.** For this criteria, Bloor evaluated how well vendors are utilizing their wider product portfolios or partner channels to deliver end-to-end streaming analytics solutions.
- 3. Edge processing and the IoT.** For this criteria, Bloor evaluated the vendor’s vision for the remote deployment and management of edge applications.
- 4. Platform performance and scalability.** For this criteria, Bloor assessed the platform’s performance in terms of throughput and latency, including scalability (scale-up and scale-out), resilience and fault tolerance.
- 5. Productivity tools and application development ease.** For this criteria, Bloor assessed the vendor’s programming language, development and test tools, including the process of transitioning changes into operational use, plus use of other libraries and solution accelerators.

While the five Bloor Research evaluation criteria differed from the Forrester criteria, the results were the same. IBM was named a champion for enterprise stream processing and cloud streaming analytics.

To support this champion designation, Bloor highlighted the maturity of IBM Streams and its features. And they observed IBM Streams customers reside in a wide range of industries, such as telecoms, government, financial services and automotive.

Bloor also pointed out that IBM Streams has flexible deployment on premises or on the cloud, and now offers integration with IBM Bluemix® for IoT application development. In addition, IBM Streams integrates with

Apache Edgent to run analytics at the edge by doing correlations across millions of devices like telematics and automobiles.

Bloor could not find fault with the set of operators IBM Streams includes for data stream processing. They also noted the performance and scalability of Streams with impressive scalability benchmarks for single server and on multi-node clusters, along with high availability, self-healing and automatic recovery.

They noticed that Streams includes its own language called Streams Processing Language (SPL) to build applications with additional support for operators built in C/C++, Java, Scala and Python. By enabling support for a variety of languages that can be deployed in the Streams runtime, it expands the range of developers that can rapidly build real-time applications.

Finally, they spotlighted available productivity support tools, such as graphical development tools, with connectors for databases, NoSQL storage platforms and messaging systems. Of particular note were the pre-built toolkits for functions such as machine learning, geospatial operators and original design manufacturer (ODM) business rules that improve programmer productivity and return on investment (ROI).

“Streams is a competent and successful product that deserves its place as one of the leaders in our report.”

Bloor Market Report: *Streaming analytics 2016*

IBM Streams in the real world

While analyst rankings are valuable when evaluating product offerings, real-world performance provides the ultimate test. And Walmart has discovered that the performance advantages of IBM Streams mean more money for its bottom line.

With 2.2 million associates worldwide, each week Walmart serves nearly 260 million customers through more than 11,500 stores under 72 banners in 28 countries, and

e-commerce sites in 11 countries. This global enterprise creates tremendous data challenges, including:

- Monitoring global logistics for supply chain accuracy
- Updating inventory in real time for up-to-date supply chain status
- Analyzing data in real time to adjust to market dynamics

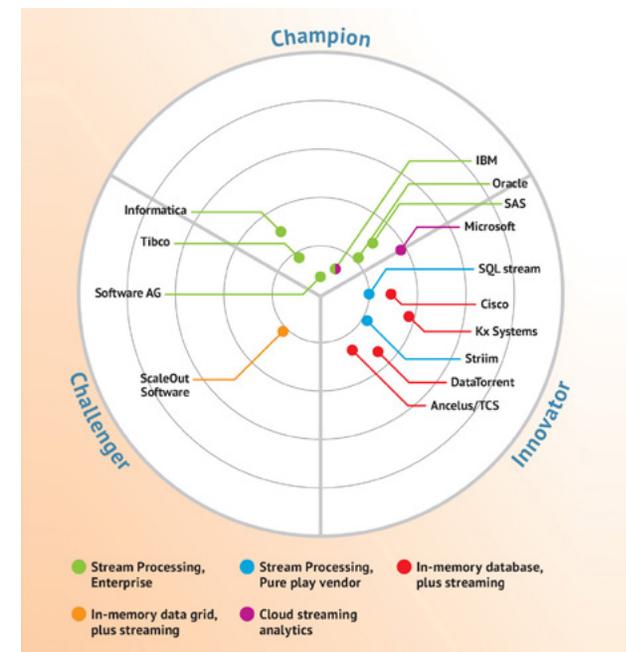


Figure 3. Bloor Research Market Report rankings

With so much at stake, streaming analytics are not optional for Walmart—they're a business requirement. And with such large volumes of data around the globe, even small performance and cost improvements due to streaming analytics can make a big difference.

To ensure its streaming analytics deployment maximized its ROI, Walmart searched for an objective measure it could use. With no comprehensive streaming benchmark available, it identified the Linear Road Benchmark for Stream Data Management Systems (SDMS) as the best tool to obtain an objective system comparison.

Linear Road streaming benchmark

The Linear Road Benchmark is an open source benchmark originally developed in 2004. It allows SDMS solutions to be compared against each other, as well as to alternative systems, such as relational databases using a toll system simulation for the motor vehicle expressway in a fictional metropolitan area.

The tolling system uses variable tolling, which uses dynamic factors such as traffic congestion and accident proximity to calculate toll charges. Specifically, Linear Road includes features like accident detection and alerts, traffic congestion measurements, toll calculations and historical queries.

Since it was written in 2004, Linear Road no longer represented the current state of streaming data systems—it was too slow for current throughput and hardware capabilities. So Walmart rewrote the original Linear Road benchmark, including the data generator and results validator, and made it publically available on [GitHub](#).

The rewritten Linear Road Benchmark still modeled a 100-mile-long by 100-mile-wide fictional metropolis with 10 horizontal expressways evenly spaced across the city. And the model was still based on these details:

- Every mile each has an exit and on-ramp.
- Each expressway has four lanes in each direction, three travel lanes and one lane for entrance and exit.
- Every vehicle emits a position report every 30 seconds.
- One accident occurs randomly on each expressway every 20 minutes, taking 10 to 20 minutes to clear.

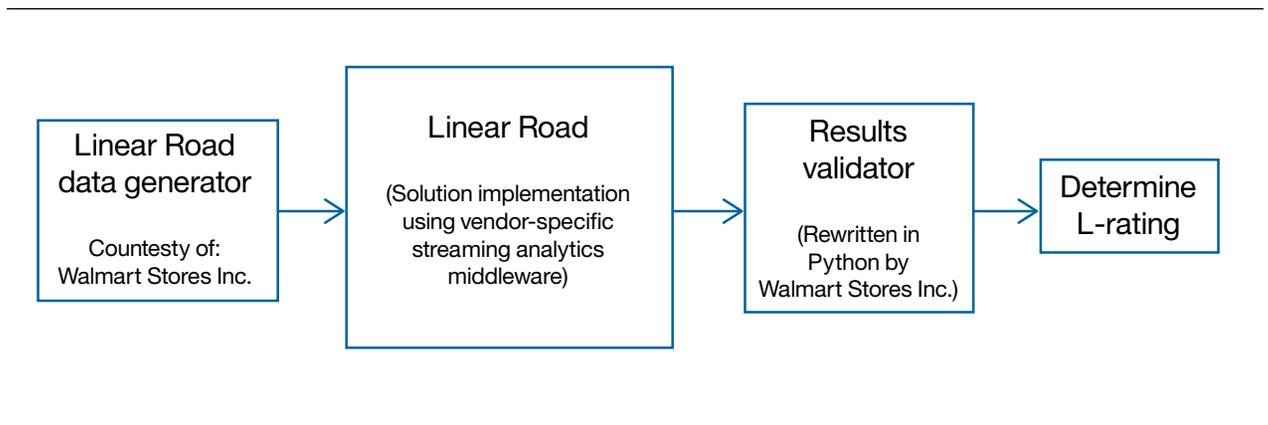


Figure 4. High-level Linear Road architecture

Further, the benchmark still required processing a set of continuous and historical events generated from the model's traffic input stream. The four notifications and queries from the vehicles include:

1. Real-time position reports—comprising 99 percent of the events
2. Daily expenditure queries for a specific day in the past 10 weeks—comprising 0.50 percent of the events
3. Historical request queries for toll account balances—comprising 0.10 percent of the events
4. Travel time prediction queries—comprising 0.40 percent of the events

The goal is to process every event emitted from all cars in the simulation in a timely manner, which varies depending upon the event. Streaming engines can then be evaluated based on how many expressways they can handle, expressed as the L-rating. The higher the L-rating the better the performance.

Real-world results

Using the updated Linear Road Benchmark, Walmart compared three solutions: IBM Streams, Apache Apex and Apache Storm. It started by using the same set of hardware constraints for each streaming engine—all operating on the Microsoft Azure cloud. Then, each engine was optimized to handle the large amount of incoming data, process it and write it to an output file.

When the results were in, one product pulled ahead of all the others. With an L-rating of 200, IBM Streams clearly outperformed the Apache Apex L-rating of 102, and left Apache Storm and its 10 L-rating in the dust.

But that upper L-rating, or number of expressways, is only a small piece of the story. In running this comparison, Walmart discovered that IBM Streams handles heavy traffic bursts extraordinarily well. One reason—IBM Streams executed 99 percent of the notification and query responses in sub-second time frames.

Figure 5 shows the full table of the Linear Road Benchmark results for IBM Streams. You can see the CPU usage dropped when the L-rating jumped from 50 to 100. This drop reflects the scaling from one node with 16 cores to four nodes and 64 cores.

In analyzing the data, a clear pattern emerges – IBM Streams shows linear scalability throughout the entire performance range. Based on the three-hour benchmark test runs, the throughput showed that IBM Streams consistently averages 100,000 events per second per node.

# of x-ways	# of cars	Entries	Memory (GB)	CPU (%)
1	278,973	19.2 million	2.2	2
2	558,726	38.5 million	4.5	4
5	1.3 million	96.3 million	10.9	7
10	2.7 million	192.5 million	22.0	11
15	4.1 million	289.7 million	33.0	16
20	5.6 million	385.2 million	43.5	20
25	6.9 million	482.0 million	54.5	26
50	14.0 million	963.1 million	109	31
100	27.6 million	1.9 billion	220	22
150	41.5 million	2.8 billion	330	33
200	55.0 million	3.8 billion	440	45

Figure 5. IBM Streams run data on Linear Road Benchmark

Just as important to Walmart was the clear performance advantage IBM Streams delivered to its bottom line. To achieve performance comparable to the five Azure nodes required by IBM Streams, Apache Storm required 100 Azure cloud nodes.² That 20 times hardware difference translates to a large annual cost savings.

Further, the design of IBM Streams allows for design, deployment and operation efficiencies as well. For example, this IBM Streams development effort took one person just 14.5 days to complete:

- A day and a half to install Linux and IBM Streams on five Azure nodes
- Two days for application design
- Eight days of iterative development
- Three days for unit testing and tuning

Some efficiency is due to the integrated development environment (IDE) in Streams that includes wizards for common tasks and a graphical programming interface for connecting operators with drag-and-drop ease. Other efficiencies are due to the automated scaling with user-defined parallelization included in Streams, which helps reduce time to develop scalable applications.



IBM Streams performed 20x better than Apache Storm and 2x better than Apache Apex.

What it means to you

As technology continues unabated, your organization's success will increasingly be defined by how well you uncover insights. Yet with the nonstop growth of data volume and velocity, insights generated from data streams can lose value by the time traditional analytics tools discover them.

So extracting actionable insights means one thing—real-time analytics allow you to take immediate actions. These actionable insights are exactly what streaming analytics helps make possible.

But not all streaming analytics solutions are equal. Neither are all solution providers.

As Walmart discovered in a head-to-head test, IBM Streams significantly outperformed the competition. And when easier deployment and better scalability are added to the performance benefits, IBM Streams can improve your bottom line.



Yet the performance, deployment and usability advantages of IBM Streams are just part of the equation. IBM also brings strategic vision and the ability to globally support your needs. The combined advantages of Streams and the IBM organization are why companies like Walmart and industry experts like Forrester and Bloor Research have concluded that IBM Streams is a leader for streaming analytics solutions.

About the author

Preetam Kumar is the Portfolio Marketing Manager for IBM Analytics where his more than 10 years of experience in Indian, US and UK markets helps clients achieve their analytics goals. Prior to joining IBM, Preetam served as a senior business development consultant for Oracle where he actively contributed to the customer alignment program. He holds a B.E. in Mechanical Engineering from KIIT University.



© Copyright IBM Corporation 2017

IBM Corporation
Sales and Distribution
Route 100
Somers, NY 10589

Produced in the United States of America
March 2017

IBM, the IBM logo, ibm.com, and Bluemix are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml.

Microsoft and Azure are trademarks of Microsoft Corporation in the United States, other countries, or both.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

The content in this document (including currency OR pricing references which exclude applicable taxes) is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. All client examples cited or described are presented as illustrations of the manner in which some clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions. Contact IBM to see what we can do for you.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

1 IDC. “The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things.” April 2014. <http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>

2 Calculated using Microsoft Azure pricing calculator with D14 v2 instance size, 16 cores, 112 GB RAM and 800 GB disk at USD 2.111 per hour, 29 January 2017. <https://azure.microsoft.com/en-us/pricing/calculator>



Please Recycle