

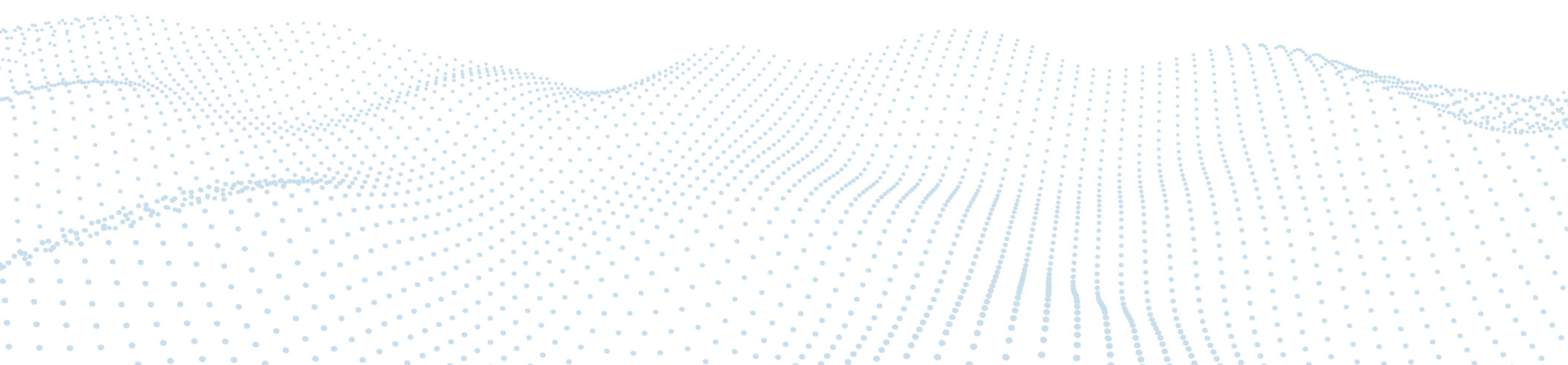


Streaming Data Simplified

Operationalizing Pipelines for Real-Time Business Decisions

Table of Contents

Introduction	1
What is a Streaming Data Pipeline?	2
Streaming Data Pipeline Benefits.....	3
Streaming Data Pipeline Fundamentals	4
How Do Streaming Data Pipelines Help Your Business?.....	5
How Streaming Data Pipelines Enable ML.....	5
Streaming Data Pipelines in the Fortune 100	6
Hewlett-Packard Enterprise Delivers Customer Value Faster.....	6
John Deere Increases Market Share.....	6
Conclusion	7



Introduction

“Real-Time” means different things to different people. For some engineers, it means sub-millisecond. To a data analyst, it might mean data updated within the last hour. But for organizations, real-time equates to receiving in-the-moment insights from data to make business-critical decisions while providing unsurpassed customer experiences.

To enable real-time insights, enterprises are investing in big data, artificial intelligence (AI), machine learning (ML), and building data pipelines that can apply ML models to data as it's being generated. Known as “streaming,” these organizations are incorporating streaming data pipelines for transformational payoffs.

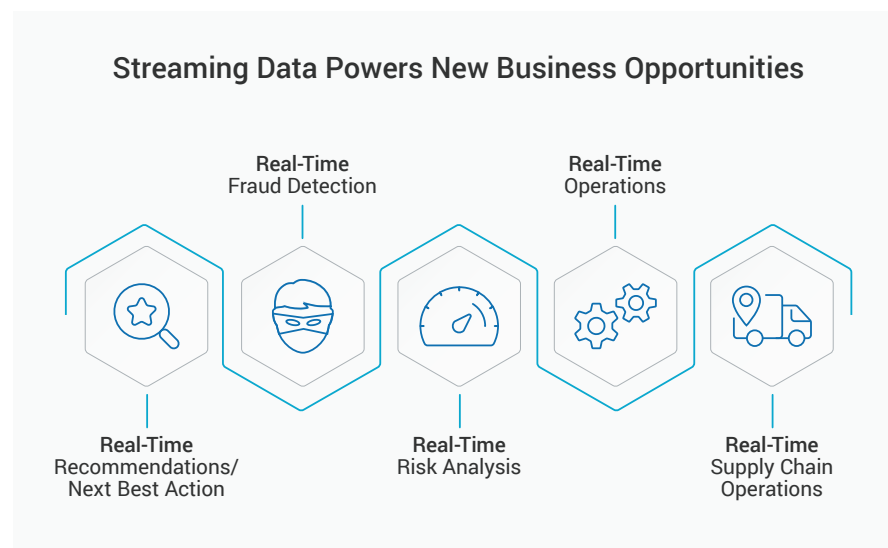
Today, the development of a reliable streaming data pipeline is constrained by complexity at one end, and fragility at the other. Simple or easy to use solutions require significant operational investment, and complex solutions that need to be inherently stable often burden teams with many of the reliability concerns that rightly belong as part of the infrastructure. The problem is meeting both of these constraints while trying not to sacrifice simplicity for reliability.

Real-Time Decisioning Done Right

The demand for availability, scalability, and resilience is forcing streaming architectures to become more like microservice architectures. Conversely, successful organizations building microservices discover that their data needs to grow with their business while their data sources are becoming more stream-like and more real-time. As a result, there is a natural unification between streaming data and microservice architectures.

Streaming data that leverages microservices can accelerate time-to-value and make it even simpler to develop, build, deploy and manage an ‘always on’ streaming data architecture. While having immediate access to impactful data is an imperative, optimal value can be achieved when ML is applied to this data. But to do this requires the ability to operationalize the disparate components, from data ingest through analysis, that processes vast amounts of data.

In this ebook, we will provide the foundation to get started on your journey to streaming data pipelines, including what this technology is, what it does, and the value it can bring to your enterprise.

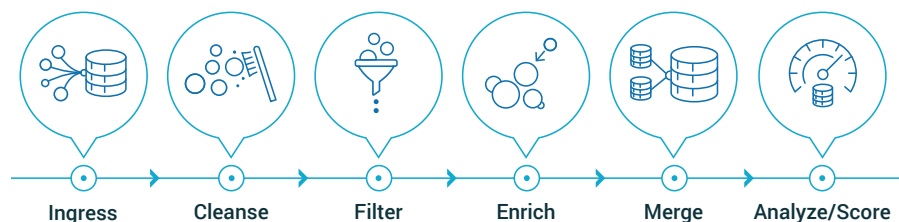


CHAPTER 1

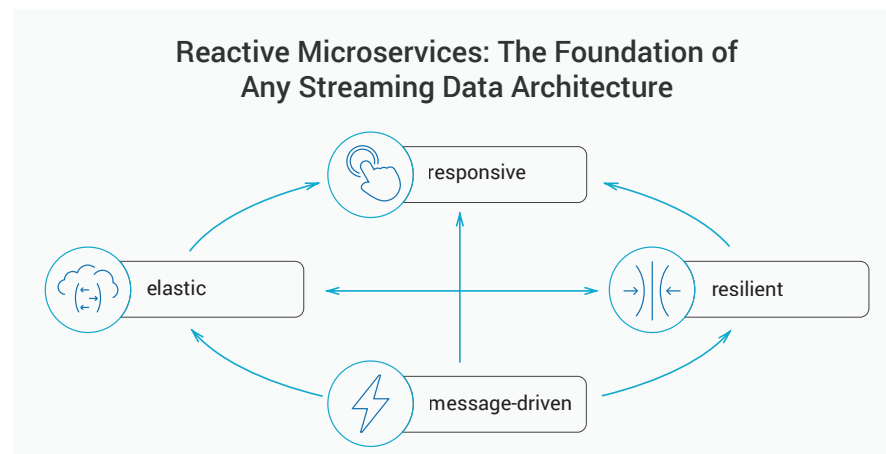
What is a Streaming Data Pipeline?

Enterprises are under increasing pressure to build real-time intelligence into their business activities, such as recommendation engines, real-time personalization, real-time risk analysis, real-time supply chain optimization, IoT operational controls, and financial services processes. This means every business can benefit from better use of their data.

Currently, regardless of where raw data may come from—for example, customer data, market data, device data, sensor data, social media feeds, application logs, or transaction logs—to extract value often requires several processing stages that may look something like this:



Linking all these stages together is referred to as a data pipeline. However, if the data is to be processed continuously as it is generated or comes into your system, then this becomes a streaming data pipeline.



A **Reactive microservice** provides the building blocks necessary for streaming data in a microservices system, while remaining very lightweight for efficiently using cloud resources. Your company can harness these features for new development, modifying or refactoring your existing applications, and creating integration points with legacy systems. Through these efficiencies, you decrease your infrastructure expenses while accelerating development productivity and time to value.

Streaming Data Pipeline Benefits

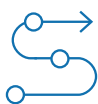
The ability to make data actionable and extract value from that data in real-time is a 'must have' organizational strategy in today's competitive landscape. By building a streaming data pipeline, your company can realize significant business benefits while maintaining a stable system:



Real-time insights. Streaming data pipelines enable organizations to access data as soon as it becomes available. This type of instant feedback helps businesses make the best decisions they can with the most up-to-date information and keep a competitive edge.



Break down data silos. Streaming data pipelines help make all enterprise application data accessible and usable – and, in turn, begins to break down the data silos that commonly exist within large environments.



Evolve business applications. As data collection from streaming data pipelines grows over time, the data can be fed into more complex processes to create next-gen applications driven by AI or ML.



Improve customer experiences. Streaming data pipelines allow organizations to gain insights into the needs of their customers. By leveraging these insights, organizations can begin to create more customized customer experiences.



Akka is an open-source, message-driven toolkit available on the Java Virtual Machine (JVM) that enables development teams to build Reactive microservices. This toolkit breaks down application logic into individual services called “actors,” or autonomous processes.

Akka is downloaded more than 70 million times a month by developers around the world who praise the technology for its ability to fail and self-heal, delivering exceptional resilience and elasticity.

Akka fully supports Kubernetes for cloud native development, and is currently in use in major Global 2000 companies including Starbucks, Nike, Hewlett-Packard, Amazon, and others.

“Akka has consistently allowed us to cut 80% of infrastructure, or increase overall application performance by 5x, when compared to the traditional systems we replaced.”

Akara Sucharitakul, Principal MTS



CHAPTER 2

Streaming Data Pipeline Fundamentals

Streaming data pipelines represent a new frontier in business technology, one that allows you to maintain a competitive advantage and analyze large amounts of information in real time. The right tools enable you to build and maintain your streaming data pipeline and ensure data accessibility across the enterprise.

To get started building a basic streaming data pipeline, you'll want to consider the below steps and tools:

1. **Select a stream processing engine** — such as Apache Spark, Apache Flink, Akka Streams, or Kafka Streams— capable of handling batch, real-time analytics, and data processing workloads
2. **Choose a streaming platform** — such as Apache Kafka — to construct the actual pipeline
3. **Cleanse, validate, and write the data** to ensure it's useful and formatted correctly
4. **Initialize the stream processing engine** and establish the streaming context (e.g., training ML models)
5. **Fetch and transform the data** from the streaming platform
6. **Manage the pipeline** to ensure everything is working as necessary

Understanding Your Stream Processing Engine Options



- Best known; large community
- Batch, mini-batch, and new low-latency streaming
- SQL and ML use cases



- Low-latency streaming
- Rich dataflow language
- Rich APIs for microservices, data sources, and sinks
- Excellent for model serving



- Low-latency streaming
- SQL and ML use cases



- Read, write Kafka topics
- Stream and table abstractions
- SQL on streams

How Do Streaming Data Pipelines Help Your Business?

Streaming data pipelines—built to get the most out of your streaming data application architecture—offer significant advantages for your IT and ML efforts, contributing directly to the success of your business.

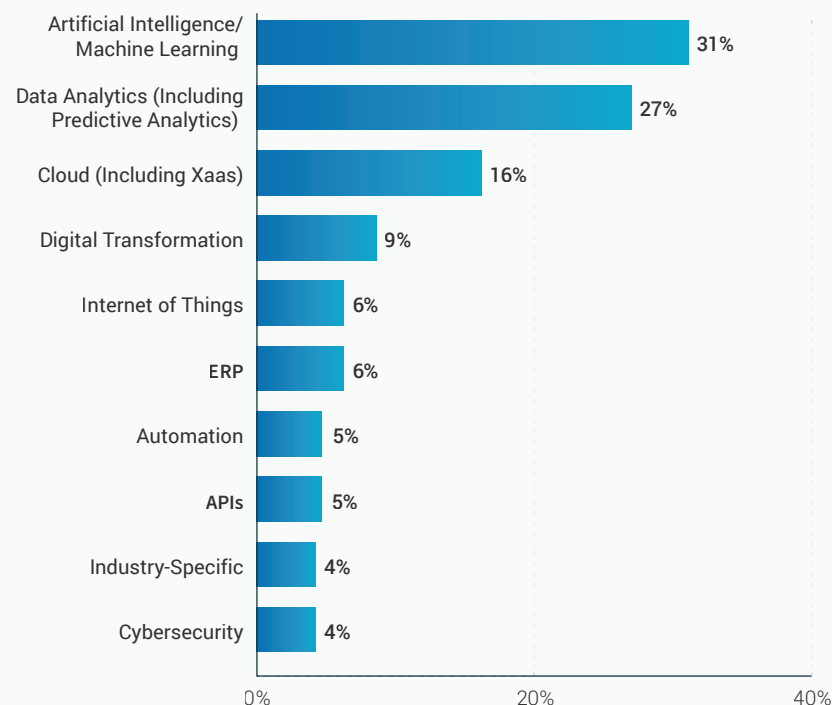
Streaming Data Pipeline Advantages for Executives	Streaming Data Pipeline Advantages for Development Teams
<ul style="list-style-type: none"> • Improve business operations • Increase customer loyalty • Optimize costs 	<ul style="list-style-type: none"> • Happier developers • Focus on writing mission-critical business logic • Build new software and features more quickly

How Streaming Data Pipelines Enable ML

Organizations invest in big data to get more insight into their business. They further invest in data science teams to build and train ML models for making better business decisions. Both are excellent steps, but the greatest value can be realized when true operationalization of data for real-time business decisions can take place. And that requires streaming data pipelines.

For context, let's say you want to determine the optimal location to open your next warehouse. Lots of data is gathered: predicted customer locations, historical buying trends, local tax breaks, regional cost of labor, weather patterns, and proximity to key links in the supply chain. This data is used to train a ML model, which takes additional time to refine before coming up with suitable options.

Top 10 “Game-Changing” Technologies for the Enterprise



n=1,055 All respondents, excluding “prefer not to answer”
 Q. Which technology area do you expect will be a game-changer for your organization? Coded open-text responses. Multiple responses allowed. Showing 10 most common responses
 Source: 2020 Gartner CIO Survey ID: 450986

CHAPTER 3

Streaming Data Pipelines in the Fortune 100

Large-scale companies are building streaming data pipelines so they can stay competitive while improving business operations. Let's explore how two Fortune 100 companies have been successful using streaming data pipelines to transform their systems and grow their bottom line.



Hewlett Packard Enterprise

Hewlett-Packard Enterprise Delivers Customer Value Faster

Hewlett Packard Enterprise (HPE) has sensors deployed in data centers all around the globe sending trillions of metrics each day to InfoSight, an industry-leading predictive analytics platform providing analytics on petabytes of telemetry data. However, to deliver value faster, InfoSight needed to evolve from its classic batch mode, big data architecture to streaming data.

HPE turned to streaming data pipelines powered by Reactive microservices for the elasticity and resilient self healing required to deliver on the promise of streaming data. The business benefits the company has experienced since implementing streaming data pipelines include:

- **Accelerating time to value** — Deliver near real-time insights from over 20 billion sensors sending trillions of metrics each day.
- **Increasing developer productivity** — Utilize simple frameworks for processing continuous application logic.
- **Deepen customer experiences** — Customers can now monitor infrastructure, predict possible problems, and learn ways to enhance performance.



John Deere Increases Market Share

John Deere's farming equipment has become incredibly complex, holding more than a 1,000 sensors that provide an immense amount of data. Their customers wanted to know how efficiently teams are operating these complex and expensive machines to ensure investment optimization. This required real-time analytics for rapid decision making.

Driving John Deere's real-time journey is an architecture that leverages streaming data pipelines. This architecture has enabled greater business success by:

- **Accelerating time to value** — Created an entirely new agriculture service for competitive advantage.
- **Increasing developer productivity** — Developers can now deliver new features more quickly and with fewer failures.
- **Deepen customer experiences** — Customers are able to optimize equipment operation for far more efficient crop yields.

Conclusion

Streaming data pipelines can introduce a significant amount of complexity for application developers, data engineers, and DevOps teams. Different processing engines are often required for different stages of the pipeline, and applying ML models to the real-time stream of data is not exactly a simple exercise.

Additionally, and most importantly, data never stops coming. Building reliable, scalable streaming systems is extremely difficult as is, not to mention potential data and system faults. Given a never ending stream of data, all the edge cases that very rarely occur will all eventually occur. Planning for failure, recovering completely and gracefully, is critical — and again, difficult.

So, yes, building streaming data pipelines and running them successfully in a production environment is hard. But despite these challenges, the business benefits — as described in this eBook — are definitely worth it.

Akka Data Pipelines has emerged as a leading technology framework to quickly construct streaming data pipelines. Using the stream processing engine of your choice, Akka Data Pipelines provides the tools that remove the headache and time associated with deploying and managing streaming data pipelines. This frees teams to focus on developing business logic and applying machine learning models that infuse intelligence into data for real-time decisioning.

Even better, Akka Platform, a set of building blocks for rapidly building, testing and deploying Reactive microservices, compliments Akka Data Pipelines for a comprehensive solution that further accelerates developer productivity, time-to-market and bottom-line growth.

What Should My Organization Do Next?

Being able to analyze and operationalize data in real-time brings other benefits such as improved customer service, optimized supply chains, enhanced manufacturing operations, and more.

And if you are not taking the steps to move to real-time, you can bet that your competitors and new market disruptors are.

If you'd like to get started with streaming data pipelines, [explore how Lightbend can support your enterprise](#) — and see first-hand how building pipelines can drive competitive advantage.



Lightbend (@Lightbend) is leading the enterprise transformation toward real-time, cloud-native applications. Lightbend Platform provides scalable, high-performance microservices frameworks and streaming engines for building data-centric systems that are optimized to run on cloud-native infrastructure. The most admired brands around the globe are transforming their businesses with Lightbend, engaging billions of users every day through software that is changing the world.

For more information, visit www.lightbend.com

