Whether Ghostbusting or Analyzing Data: Cross the Streams

By Pete Goddard | October 28, 2021

Harnessing data streams — joining both batch and real-time events — empowers data scientists and analysts to address sophisticated problems.

Individual streams provide data related to a particular dimension — the price of a stock, the order of a customer, the metric of a device. Analytics and applications can be served by a single stream of data, but uses are narrow and local.

Crossing streams unveils grander possibilities, ones filled with history, context, and related signals. When our Ghostbuster heroes (Venkman and the gang) needed to rise to the challenge (and defeat Mr. Stay Puft), they joined forces—and streams! The whole was greater than the sum of the parts.

In our community, data scientists, analysts, and developers are similarly called to action. Harnessing data streams — joining both batch and real-time events — empowers you to address sophisticated problems. And, as

CDInsights

Report: Cloud Data Management Lags Cloud Adoption

6 Ways Supercloud Might Impact Emerging Cloud-Computing Trends

National Labs Provide a Peek into the Future of Data Sharing

Data Sharing Key To Smart City Project Success

First Steps toward Leveraging Enterprise ChatGPT

with Venkman, sometimes you need others to bring their gear and help. Here are four vital components to making the crossing of streams successful:

1) Bring together data, use cases, and people.

Accelerating innovation, maximizing efficiency, and providing flexibility are established priorities for sophisticated data systems. A nimble, evolving software backbone realizes these goals. Open-source core components provide the long-term agility and interoperability paramount for success.

Tools evolve, and sometimes you need to use that new ghost trap.

2) Future-proof your data stack with open-source formats.

Data portability has long been a sacred requirement for enterprise data teams. Walled gardens create future debt, and vendor lock-in has an unspoken long-term cost, one often paid in business drag. **Store data using open formats.**

CSV and JSON have been big for years, with Avro, Protobuffs, Parquet, Orc, and others recently gaining popularity. They have respective reasons to exist, but each is principled on the delivery of structured data to a plethora of independent systems, agnostic to and oblivious of the computer science downstream.

As the magnitude of data has scaled and the related financial and latency cost of moving data has compounded, the concept of open data now includes in-memory formats, not just the kind that persisted on disk. It is now often unacceptable to require data to be copied, moved, serialized, or translated in any way. In particular, Apache Arrow's significant community benefits from its ability to serve in-memory data to a range of data processing libraries across many languages with minimal overhead, zero-copy reads, and fast access at scale.

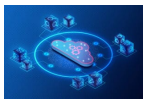
But let's remember, in Ghostbusters, the data was just the start of the adventure.

3) Make joining real-time and static data a fundamental requirement.

A modern data engine must bring together data from a variety of sources. The jargon of warehouse, lake, and the centaur-like lakehouse are now

Connect with us

Spotlight


Data as a Product Spurs the Need for a Data Development Platform
July 28, 2023


Bringing SAS Analytics Programs into the World of Modern Development
April 3, 2023

Data Management and AI: Creating Value at the Edge
March 15, 2023

When Clouds Deliver Snow, a Cloud Database Helps Improve Plowing Efforts
March 1, 2023

Content Hubs

Smart Manufacturing for Automotive

Center for Data Pipeline Automation

common imagery. However, the growing popularity of event streams is a not-so-quiet canary suggesting static data is no longer the whole story.

Data changes. Modern workloads live in a state of flux. **Real-time data matters.**

Data engines and processing libraries must be architected to address and move fluidly between real-time and static data workloads. "Continuous intelligence" is a trendy phrase for systems that combine the context of history with the event signals of the moment. Modern data systems should be built to process real-time data, event streams, and other updates as a first-class competency. These should be core strengths, not add-ons, not afterthoughts.

After all, as we learned in Ghostbusters, Gatekeepers and Key Masters are a lot less powerful until they are joined together.

4) Always put the user first.

Today's data users have a variety of skills, tools, workflows, and priorities. Coalescing a team around a shared platform serves the individual while energizing the team. Data systems that maximize individuals' efficiency and foster collaboration drive business value.

Open data software lights the way. The intriguing mix of cooperation and competition in open projects yields an unrivaled pace of progress and ingenuity. Organized to encourage interoperability, community development promises enhancements, integrations, and user experience upgrades. Popular paths become paved roads. Such systems make users an army of one while supporting the codependent work product required for any even moderately complex use case.

After all, one proton pack is powerful, but four working together is invincible.

I ain't 'fraid of no ghost.

Tagged Real time data streaming data

Improving Service and Profits With Connected Products

Center for Automated Integration

Center for Edge Computing and 5G

Center for Observability and AIOps

Continuous Intelligence: Insights

Event-Driven Architecture for the Cloud

DSP Resource Pages

Enabling Real-Time Action with Stream Processing

Activate your Mainframe Data for Cloud Analytics

Multi-dimensional data observability

Simplifying database access from Kubernetes

Real-time Location Intelligence

Today's Low Code Integration Platform

Real time data pipe-lines with Apache Pulsar

About Pete Goddard

Pete Goddard is the CEO and co-founder of Deephaven Data Labs, a data company building software for modern data teams. After founding quantitative trading company Walleye Capital in 2005, Pete and his engineering team were searching for ways to help quants, data scientists, developers, and portfolio managers discover and evolve strategies and signals more quickly. After witnessing how Walleye benefited from the solution they built, Pete took those engineers, the data system, and its related IP out of Walleye and formed Deephaven as an independent company.

View all posts by Pete Goddard →

Recommended Articles

Making Remote Work More Effective with Enterprise Search

Modernizing Core Banking and Insurance Services

Why Your Digital Twin Will Be Essential in the Metaverse

What Autonomous Vehicles Can Learn from IoT about Real-Time Design

Leave a Reply

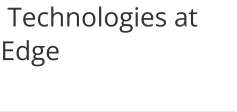
Your email address will not be published. Required fields are marked *

Comment *

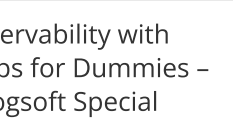
Name *

Email *

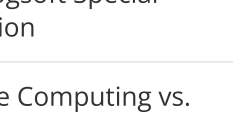
eBook, Primer, Videos and Podcasts



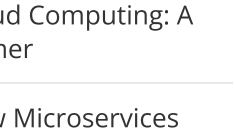
Dell Technologies at the Edge



Observability with AIOps for Dummies – Moogsoft Special Edition

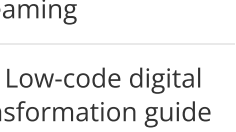


Edge Computing vs. Cloud Computing: A Primer

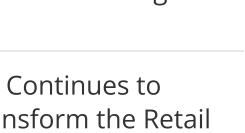


How Microservices Developers in Financial Services Use Streaming

The Low-code digital transformation guide



IoT Continues to Transform the Retail Experience in 2021



Digital Twin: Closing the Loop from Operations to Design

Recent Articles

How AI Helps Businesses Fight Fraud
August 1, 2023

Using Photonic Neurons to Improve Neural Networks
August 1, 2023

It has Begun: AI-Based Attacks Against Businesses
July 31, 2023

AI Revolution To Hit Law, Medicine, Finance Hardest
July 31, 2023

Website

Save my name, email, and website in this browser for the next time I comment.

Post Comment

Real-time Analytics News for Week Ending July 29

July 30, 2023

What's Trending

5G AI analytics Artifical intelligence artificial intelligence automation autonomous vehicles big data blockchain CIhub cloud cloud-native cloud database Continuous intelligence COVID-19 customer experience cybersecurity cyber security data data management data security DevOps digital transformation digital twin digital twins edge computing healthcare ibm IIoT industrial IoT Industry 4.0 Internet of Things iot IoT security Kubernetes machine learning Observability predictive analytics predictive maintenance real-time analytics real-time data robotic process automation smart cities streaming analytics supply chain

Featured Resources

7 Data Lake Best Practices for Effective Data Management

Follow these best practices for data lake management to ensure your organization can make the most of your investment.

What's Changing Faster? Data Pipeline Tech or the Role of the Data Scientist?

The need for automated data pipelines is clear. What role will data scientists play in bringing them about?

Taming MLOps: Accommodating the Needs of Different Developers

Developing an enterprise-ready application that is based on machine learning requires multiple types of developers.

Reducing Cloud Spend During Economic Uncertainty

Cloud optimization could offer the best method for reducing costs according to a new report.