Feeding the Monster: The Challenge of Streaming Analytics

For the first time in the seventy-year history of data processing, we are at the cusp of a sea change. For all those decades, we collected data first and analyzed it later – often with multiple intermediate processing steps thrown in for good measure.

The most obvious aspect of this shift is the move from delayed to real-time analytics. We don't have to wait to summarize or transform the data any more. Real-time insight is now at our fingertips.

There is a more subtle change afoot, however. That entire lifetime's worth of data processing meant processing data *sets*: set quantities of data that we would collect and then analyze.

Even as the conversation shifted to big data a few years ago, we still thought in terms of data sets – increasingly large data sets to be sure, but data sets nevertheless.

Today, in contrast, we're increasingly focusing on the streams of data themselves – especially high-volume streams. For example, imagine how you would go about gathering insight from all the tweets on Twitter. You could theoretically fill a large Hadoop cluster with tweets and crunch those – but the tweets just keep on coming. The challenge with such data streams is gathering insight from the stream *in motion*.

The field of streaming analytics takes as its starting point not simply large data sets, but this never-ending, 24 x 7 fire hose of data. All of a sudden, making sure we have no bottlenecks simply becomes the price of admission.

We can't even begin to deal with streaming data if we have to pause even for a moment to move it, store it, process it, or analyze it. The fire hose just keeps on streaming.

The advantages of streaming analytics are profound. Fraud detection and prevention, dynamic product pricing, Internet-of-Things (IoT) data analysis, electronic trading, customer promotion triggering, and compliance monitoring are some of the early examples of the power of streaming analytics.

Today's technology is finally piecing together the end-to-end components that make streaming analytics a reality. Tools that can process and analyze high-volume data streams are maturing rapidly, and new entrants – both commercial and open source – appear on the scene with surprising regularity.

For many enterprises, however, the challenge is less about processing and analyzing data streams – it's about collecting them in the first place. After all, the streams coming from web scale companies like Twitter and Facebook swamp the existing streams from more traditional companies.

There are plenty of examples of enterprises who do generate high volume data streams. Perhaps the most notable is General Electric, who has turned every machine they produce, from MRI machines in hospitals to locomotives to wind turbines, into a source of streaming data. Airplane manufacturers have done the same with their products.

Other enterprises, in contrast, struggle with the notion of high-volume data streams – not because they don't have the data, or because they aren't producing them fast enough. Rather, the enterprise challenge is often that the data sources are scattered about the organization, with no real-time way of bringing them together.

There may be transactional data from one system, streaming server log files from another, and miscellaneous data streams, from building security systems to factory and warehouse equipment. True, the Internet of Things may account for some portion of this streaming information, but the lion's share are likely to come from more traditional sources.

Business stakeholders are increasingly realizing that there is value in such data – but only if they can properly coordinate and integrate those streams in order to support the streaming analytics that promises to provide the deep insight into the business that is the promise of big data.

Pouring the data into a data lake won't address this problem. Traditional data integration technologies aren't up to the task either. Instead, enterprises require the ability to integrate data streams on the fly in an inherently scalable, stateless manner.

WE NEED TO APPLY CORE CLOUD ARCHITECTURAL PRINCIPLES TO THE DATA INTEGRATION WORLD.

In other words, we need to take a page out of the cloud computing playbook and apply core cloud architectural principles to the data integration world. Few vendors are able to accomplish this feat at scale.

SnapLogic's Ultra Pipelines are among the short list of products that bring the cloud to data integration in a way that will support streaming analytics. SnapLogic helps bridge the two worlds of traditional data sources and cutting edge analytics to help feed the monster that is streaming analytics.

Don't let the fire hose of data knock you over. Remove bottlenecks and squeeze insight out of diverse enterprise data streams with data integration that is built to scale.

SnapLogic is an Intellyx client, but Intellyx retains full editorial control of this article. At the time of writing, no other organizations mentioned in this article are Intellyx clients. Intellyx retains full editorial control over the content of this article. Image credit: David Moss.