INFORMATION MANAGEMENT HANDBOOK

MARCH 2017

FOTOLIA

Companies tap streaming analytics tools for real-time business view

TechTarget

In this handbook:

Data streaming applications set a faster pace on analytics efforts

CRAIG STEDMAN

The big data analytics team at Comcast Corp. wants to see a lot of information in real time. The list includes data on the locations of the media and entertainment company's customer service trucks, phone calls to its call center, the performance of set-top boxes and aggregate TV viewing records.

"We want to be able to fix issues before customers notice them," said Kiran Muglurmath, Comcast's executive director of data science and big data analytics, during a session at Strata + Hadoop World 2016 in New York last September. To make that possible, Comcast is one of a growing number of organizations expanding their big data architectures to support data streaming applications.

What's driving such investments is an increased need for decision-making speed -- and for more clarity amid the burgeoning big data clutter. "The operational side of the business is being flooded by data. That's changing

In this handbook:

the way people manage their businesses," said Andrew Cardno, CTO and co-founder of operational intelligence software developer VizExplorer.

But not all of the data streaming into analytics systems is golden. "There's a lot of noise in streaming event data," said Mark Madsen, president of consultancy Third Nature Inc. Both he and Cardno spoke at the 2017 TDWI Leadership Summit in Las Vegas in February. Madsen noted that the data collection process with Hadoop, Spark and other big data technologies "is so much faster now, but it's not necessarily better," especially if application developers don't focus on data accuracy and consistency.

Madsen's closing advice for companies creating data streaming applications was succinct: "Manage your data, or it will manage you." This handbook offers further insight on how to make streaming analytics initiatives pay off, with more user examples and information on stream processing tools.

In this handbook:

Real-time streaming analytics systems need help from message brokers

DAVID LOSHIN

One of the growing uses for big data platforms is capturing streams of data that are continuously ingested, processed, stored and analyzed. Real-time streaming analytics provides immediate visibility into business activities and feeds operational reporting, which is particularly beneficial for organizations that can act quickly on incoming data as events unfold.

A common example is manufacturers that have equipped various plant-floor machines with embedded sensors that measure aspects of the operating environment and communicate those measurements in streams of data, increasingly via the internet of things (IoT). These data streams are fed to a central system and analyzed to develop predictive maintenance models that can identify impending equipment failures and drive pre-emptive replacements of at-risk parts, thereby reducing unplanned downtime.

Delivery and logistics services companies are another case in point. Many

In this handbook:

capture myriad streams of operational metrics on their trucks from in-vehicle computers and sensors, including data on fuel use, speed and acceleration, air intake and tire pressure. Combined with GPS-based location data that's also collected in streams and transmitted via IoT connections, the metrics can be used to assess ways to reduce fuel consumption and improve driving habits as well as to speed delivery times.

Electrical utilities have deployed monitoring devices across their power grids to transmit data about energy consumption, network stresses and potential risks of equipment failure. The data streams can be analyzed to look for ways to balance delivery of electricity, identify selected "hot" devices that could be powered down during times of high electrical usage and predict the types of parts that repair crews might need before leaving the garage.

NOT A SMOOTH FLOW IN DATA STREAMS

All of those examples share some common characteristics. In each case, there are multiple sources producing data and streaming it independently, sometimes combining streams from numerous devices to create a single logical stream. That requires a logical "broker" at the local level to meld the

In this handbook:

data and then transmit it to a centralized location.

However, that still leaves a variety of incoming data streams, and users may want to independently ingest each one into a real-time analytics system -- even to the point of differentiating between the actual data sources within a logical data feed. In addition, data from the different streams is likely to be filtered, processed and stored in asynchronous ways.

Under these circumstances, it begins to become clear that deploying big data platforms to ingest, process and analyze data streams only partially addresses the overall challenges of enabling real-time streaming analytics. Another big hurdle is organizing the methods by which streaming data is forwarded to an analytics system in a way that preserves the integrity of the operational events generating the data. IT and analytics teams need to ensure that the process of combining data streams and forwarding them maintains the order in which sensor readings, alerts and other data points are created.

That's even more challenging when operating in a distributed environment, especially when the data streams are interweaved like in a sequence of events among collaborating processes on different machines. The need to keep

things properly coordinated creates prerequisites for an overarching broker mechanism that can manage the queuing and transmission of data in all of the incoming streams.

DATA-STREAM MANAGEMENT FEATURES

Such a broker must be able to oversee the organization of streaming data by originating source, combine different streams while preserving the order of events and maintain data consistency across sources. At the same time, it has to transmit the data streams without causing any significant delays in the desired real-time delivery. And it must provide fault tolerance with assurances of recovery in the event of a failure in the data streaming environment.

Apache Kafka has emerged as the most prominent example of a fault-tolerant message broker and queuing system in the big data ecosystem. Kafka, which was created at LinkedIn and released as an open source technology, works with Spark Streaming, Storm, Samza, Flink and other stream processing platforms, as well as HBase, Hadoop's companion database. It acts as a clearinghouse for real-time message streams, providing a combination of scalability and reliability features to help address the need for high

In this handbook:

performance in streaming analytics applications involving large volumes of data.

Kafka uses a publish-and-subscribe messaging format to transmit data streams from source to target systems. Messages generated in Kafka are persisted on disk and replicated across different nodes in the server cluster that the software runs on. Because the data is replicated, multiple subscribers to different data streams can be supported simultaneously; replication also allows the tool to balance workloads across the cluster to maintain performance and data availability in the event of a node failure.

Other open source message broker technologies are also available -- RabbitMQ and ActiveMQ, for example. And as more companies recognize the potential business value waiting to be tapped in data streams, more of them likely will also see the need to deploy Kafka or another messaging system to support real-time streaming analytics applications.

In this handbook:

Users look to real-time streaming to speed up big data analytics

CRAIG STEDMAN

NEW YORK -- For more organizations, there's no time like the present to process and analyze the information flowing into their big data systems. And IT vendors increasingly are releasing technologies that facilitate the real-time streaming analytics process.

Comcast Corp. is among the real-time vanguard. The TV and movie conglomerate is on the verge of expanding a Hadoop cluster used by its data science team from 300 compute nodes to 480. In addition, Comcast plans to upgrade the system to include Apache Kudu, an open source data store designed for use in real-time analytics applications involving streaming data that's updated frequently.

"For us, the update ability is a very big thing," said Kiran Muglurmath, executive director of data science and big data analytics at the Philadelphia-based company. The Hadoop cluster, set up earlier this year, already contains more

than a petabyte of information -- for example, data collected from set-top boxes on the TV viewing activities of Comcast customers and the operations of the boxes themselves. But Muglurmath's team needs to keep the data as up-to-date as possible for effective analysis, which means updating individual records via table scans as new information comes in.

Sridhar Alla, director of big data architecture at Comcast, said doing so takes "an immense amount of time" in the Hadoop Distributed File System (HDFS) and its companion HBase database -- too long to be feasible at petabyte scale. Kudu, on the other hand, has significantly accelerated the process in a proof-of-concept project over the past three months. In one test, for example, it scanned more than two million rows of data per second. "It's writing the data as fast as the disks can handle," Alla said during a session at Strata + Hadoop World 2016 here this week.

REAL-TIME WAITING GAME COMES TO AN END

The Kudu technology was created last year by Hadoop vendor Cloudera Inc. and then open sourced. The Apache Software Foundation last week released Kudu 1.0.0, the first production version -- a step that Comcast was waiting for

In this handbook:

before going live with its Kudu deployment.

The expansion of the Cloudera-based Hadoop cluster should be completed by the end of October, Muglurmath said after the conference session. Kudu will be configured on all of the compute nodes along with HDFS, which will continue to be used to store other types of data. The data science team also plans to use Impala, a SQL-on-Hadoop query engine developed by Cloudera, to join together data from HDFS and Kudu for analysis.

Dell EMC, the data storage unit of IT vendor Dell Technologies, is also going down the real-time streaming path to support its internal analytics efforts.

The IT team is using the Spark processing engine and other data ingestion tools to funnel real-time data on interactions with customers into a combination of databases -- Cassandra, GemFire, MemSQL and PostgreSQL. Automated algorithms are then run against the data to generate up-to-the-minute customer experience scores that help guide Dell EMC's salesforce in selling tech-support subscription renewals, said Darryl Smith, chief data platform architect at the Hopkinton, Mass.-based organization.

The customer interaction data is also fed into a Hadoop data lake, but that's for

In this handbook:

longer-term customer profiling and trend analysis. For the customer scoring application, "you couldn't just throw all the data in Hadoop and say 'Go at it' [to the sales reps]," Smith said. "It's a different thing to take real-time data and do actionable analytics on it."

That does mean the same data is being processed and stored in different locations within Dell EMC's big data architecture, but Smith doesn't see that as a bad thing. "And it's not just because I work for a storage company," he joked. "If you're going to get value out of the data, you're going to need to store it in multiple places, because you're going to consume it in different ways."

One of the real-time streaming processes adopted by Dell EMC uses the open source Kafka message queueing tool to push data into MemSQL, an in-memory database designed for real-time applications. Vendor MemSQL Inc. this week released a version 5.5 update that incorporates the Kafka connectivity into a feature for creating data pipelines with exactly-once semantics -- meaning that data transmissions are processed only once, with guaranteed delivery and no data loss along the way. Smith said such a guarantee is "absolutely critical" for the kind of real-time analytics Dell EMC is looking to do.

In this handbook:

LIVING WITH SOME REAL-TIME DATA LOSS

Guaranteed data delivery isn't a necessity for eBay Inc., though. The online auction and e-commerce company uses Pulsar, an open source stream processing and analytics technology it created, to analyze data on user activities in order to drive personalization of the eBay website for individual visitors. In creating and expanding the real-time architecture over the past three years, eBay's IT team decided it didn't have to spend extra development money to build a delivery guarantee into the data pipeline.

"For our use cases, we can afford to lose a little bit of the data," said Tony Ng, director of engineering for user behavioranalytics and other data services at eBay. But Ng's team does have to keep on its toes as data flows in. For example, one of the goals is to detect bots on the site and separate out the activity data they generate so it doesn't skew the personalization process for real users. That requires frequent updates to the bot-detection rules built into eBay's analytics algorithms, Ng said.

The San Jose, Calif., company's real-time streaming setup also includes Kafka as a transport mechanism, plus several other open source technologies

In this handbook:

-- Storm, Kylin and Druid -- for processing and storing data. Ng noted that the streaming operations are a lot different from the batch data loading eBay does into its Hadoop clusters and Teradata data warehouse for other analytics uses.

"There are some constraints on how much processing you can do on the data," he said. It is eventually cleaned up and consolidated in batch mode for downstream analytics applications -- "but the things that need to be real time, we want to keep real time."

Putting together a real-time data streaming and analytics architecture is a complicated process in and of itself, said Mark Madsen, president of data management and analytics consultancy Third Nature Inc. in Portland, Ore.

Users can also tap a variety of other streaming technologies -- for example, Spark's Spark Streaming module and Apache Flink, an upstart alternative to Spark that was released in a commercial version this month by lead developer Data Artisans GmbH. But a lot of assembly is typically required to combine different tools into a functional platform. "It's a build-to-order problem," Madsen said. "[Individual IT vendors] carve out a piece of the problem, but it's hard for them to carve out the whole problem."

In this handbook:

Drizzle on tap to spur Spark Streaming architecture

JACK VAUGHAN

Innovation in Spark Streaming architecture continued apace last week as Spark originator Databricks discussed an upcoming add-on expected to reduce streaming latency.

Based on ongoing work by a lab at the University of California, Berkeley, elements of what is being called the Drizzle framework are expected to become part of Apache Spark later this year, according to the company.

The anticipated streaming update is part of Databricks' larger efforts to provide a platform for broad new analytics uses. Drizzle is intended to help promote users' moves to so-called Lambda architectures that combine batch and real-time data processing approaches.

In this handbook:

SPARK TRENDING NOW AT NETFLIX

The move to embrace both batch and real-time processing isn't an easy one, even for fast-flying web companies. But it is a natural step, according to Shriya Arora, a senior data engineer at Netflix.

Arora is part of a Netflix team that employs Spark processing and streaming to transform and push data to data scientists who develop algorithms that personalize the company's movie recommendations to subscribers. As Netflix converts some applications from batch to real time, she's working to fine-tune Spark Streaming to ensure there are monitoring alerts that warn when streaming jobs may fail.

"Streaming is better than having long-running jobs, but it comes at a cost. For example, streaming failures have to be addressed immediately. If an application is down too long, you run into data loss," she told an audience at last week's Spark Summit East 2017 in Boston.

In this handbook:

REAL TIME MEANS 'WHY WAIT?'

The real-time effort is worthwhile, however, because it can better align Netflix's movie recommendations with the immediate interests of customers. "Trending now" viewing choices, for example, can be more completely up to date, Arora said. "Why wait 24 hours when you can pick up the new information in an hour?"

But the Spark Streaming architecture today doesn't support pure event streaming -- it still has roots in a "micro-batching" formula that rapidly processes small batches of data. So, there are cases where time-sensitive applications might better opt for streaming as supported by alternative frameworks such as Flink or Storm, Arora said.

Such use cases are a prime target for Drizzle, a project within the UC Berkeley RISELab -- itself a descendent of the AMPLab project that begat Apache Spark. [RISE stands for Real-time Intelligence with Secure Execution.]

Drizzle's goal is to unify record-at-a-time streaming with micro-batch models, and is in some part an answer to Flink, an emerging streaming architecture that has shown performance benefits over present Spark Streaming.

In this handbook:

HEARING FLINK STEPS?

As he discussed Drizzle in a Spark Summit keynote, Ion Stoica didn't try to cover up Spark Streaming architecture's present latency shortcomings in streaming versus Apache Flink. He said Drizzle is intended to reduce Spark Streaming's performance latency by about 10 times.

Stoica is executive chairman and a co-founder of Databricks, and is also a professor of computer science at UC Berkeley and a part of the RISELab. In graphs, he showed Spark trailing Apache Flink by hundreds of milliseconds in handling event throughput.

He also showed data in which early versions of Drizzle and a companion Drizzle-Opt execution engine slightly improve upon present Apache Flink performance. While details were sparse, Drizzle architecture as depicted on the RISELab's website is meant to "decouple execution granularity from coordination granularity" for workloads on clusters.

In an interview, Spark inventor Matei Zaharia, who is CTO at Databricks and another co-founder -- as well as Stoica's former grad student -- said parts of Drizzle would likely appear in Apache Spark during the third quarter of 2017.

In this handbook:

PURSUING A UNIFIED MODEL

Both Stoica and Zaharia emphasized that recent advances in streaming technology for Spark, including a Structured Streaming engine and API added as part of Spark 2.0 last year, have focused on enabling a more cohesive approach for programmers that combine real-time and batch data processing on a single platform. They positioned Spark overall as a unified approach to diverse data management and analytical needs that include ETL, machine learning and SQL querying, as well as streaming.

"We think of Spark as the infrastructure for machine learning, which itself is really a small part of the entire workflow," Stoica said. "You have to clean the data, and transform it. Then, at the end, when it is curated, you apply machine learning algorithms on top."

This unified approach has merit, according to a machine learning user at a marketing analytics firm who attended the Boston event.

"Previous to our use of Spark, we had ETL, machine learning and other analytics processes, and they were all on different software stacks," said Saket Mengle, senior principal data scientist at Boston-based DataXu Inc.

In this handbook:

"Spark allows us to put this on one stack. It is something you have to tweak, but uniformity is good."

SPARK IN CONTEXT

Improvements to Spark Streaming should be viewed in the context of Spark's overall analytical adoption, said one industry analyst on hand at the conference.

"Spark's long-term appeal has been as an ensemble of analytical approaches, and its ability to address a variety of workloads," said Doug Henschen, a principal analyst at Constellation Research Inc.

In a blog post following the conference, Henschen remarked that Spark was progressing more quickly than was predecessor Hadoop at a comparable stage of development, and that it promises "wider hands-on use" by a variety of developers and data scientists.

One measure of Spark's progress is its adoption by vendors beyond Databricks, he said. In fact, the open source version, Apache Spark, is offered by traditional enterprise players like IBM and Oracle, as well as Hadoop

In this handbook:

distribution providers Cloudera, Hortonworks and MapR.

It's noteworthy, too, that Spark is offered on the cloud by the likes of Amazon, Google, Microsoft and others. So far, Databricks has focused its efforts on providing cloud services, which is where its new approach to streaming will likely first be tested.