Next-generation big data analytics tools will make sense of streaming data in real time

January 28 2019

Elke Rundensteiner, right, professor of computer science at Worcester Polytechnic Institute (WPI), and PhD student Allison Rozet, stand beside an autonomous vehicle testbed used in research at WPI. The analytics tools Rundensteiner and Rozet are developing could make driverless cars safer by analyzing data streaming from vehicles in real time. Credit: Worcester Polytechnic Institute

A new big data analytic tool being developed by computer scientists at Worcester Polytechnic Institute (WPI) will help businesses make sense, in real time, of the deluge of data that streams at them like water from a fire hose.

With a three-year, $499,753 grant from the National Science Foundation, Elke Rundensteiner, professor of computer science and director of WPI's Data Science Program, is leading a team of computer science and data science students that is building a next-generation event trend analysis tool known as SETA (Scalable Event Trend Analytics). This open-source software will be used not just to find patterns in real-time, high-volume data streams ("data in motion"), but to analyze those patterns and make sense of them on the fly for just-in-time decision making.

SETA could enable large businesses, social media sites, fraud detection centers, autonomous vehicle networks, governments, and other users to harness the continuous flow of big data as it streams in and transform it into actionable insights that could allow them to be increasingly responsive and competitive. "In a world where big data is continuously accelerating in volume and velocity, real-time streaming data analysis has become increasingly critical," said Rundensteiner, an internationally recognized expert in scalable data stream processing.

Event processing is a way to track and analyze incoming streams of information, such as online purchases, the rise and fall of a stock price, the length of time users remain on a website, or whether healthcare workers wash their hands before entering patients' rooms. It's all about flagging important events in the incoming data, so an organization can respond to them in real time. SETA will be able to handle complex queries and analytics, while providing users summarized insights cheaper and faster than is currently possible.

Most existing data analysis tools are not designed to work with streaming data, Rundensteiner noted. Instead, information must be stored in a static database before it can be analyzed, introducing a delay that might prevent the fast detection, for example, of the start of an infectious disease outbreak in a hospital. Rundensteiner's tools operate on the data as it is being generated, allowing even complex patterns to be spotted in real time, so critical decisions can be made quickly.

"Data streams are increasing at a dramatic rate, overwhelming businesses that can't make sense of their data in real time," Rundensteiner said. "By finding ways to handle these live streams, we are breaking new ground in data analytics. You could stick all this big data into a static database and look at it later, but if you want to catch a fraudulent credit card purchase as it's happening or alert a network of autonomous cars about an accident ahead, you need to analyze that information as it's streaming in at the rate of tens of thousands of pieces of data per microsecond."

With the new award, Rundensteiner will build upon her previous NSF-sponsored research in event stream analytics, which focused on finding patterns in streaming data. That work (in collaboration with former Ph.D. students, Olga Poppe, a research scientist at Microsoft Gray Systems Lab, Chuan Lei, a research staff member at IBM Almaden Research Center, and Di Wang, a research scientist at Facebook), produced analytics tools that enabled users to query a data stream for relatively simple event sequences. But if the software found many instances of the same or similar sequences and displayed them all, the user would often become overwhelmed and miss the significant patterns or the overall trends across patterns.

Rather than displaying detected sequences one by one, the new tool Rundensteiner is developing will aggregate those patterns and show the user how many times each occurs. "By showing a spike of abnormal activity, the system lets you very quickly see what is going on," she said.

"Sometimes I'm more interested in the deviation from the typical count of patterns because then I instantly know if something abnormal is happening. If one autonomous car is swerving, that might mean nothing. But if a thousand cars on the same stretch of road all exhibit deviating behavior, then something real is happening. You can then dig deeper into that particular subset of data to explore this unexpected behavior."

Developing the tools to dig deeper into these pattern aggregates is another element of the research on SETA. Rundensteiner wants to empower users to look for far more sophisticated patterns. For example, while her previous tool could be used to look for a sequence of a fixed length (say, instances of a vehicle activating the brakes, swerving, and then stopping), she wants to make it possible, with a single simple stream query, to spot sequences involving an unbounded number of instances (a car swerving an unknown number of times, braking repeatedly, and then coming to a stop, for example). While the number of potential matches to such a query could grow exponentially due to the complexity of the query language, the results promise to be more useful, she said.

To create new event trends analytics tools, Rundensteiner must first design a new query language, which is used to find and retrieve patterns in the data. By allowing users to search for more complicated patterns, the new language will make the tool significantly easier to use. She is also building a new "query engine" to process these sophisticated queries and find the requested patterns or events. A distributed engine, it will run on multiple servers across a cloud network, dramatically increasing its speed.

"Building that engine is a key part of the project," she said. "Traditionally, an engine might generate all the answers to a query, store them, and then start counting them. That's too time-consuming and expensive. Current technology might take hours, or even longer, to process a complicated query. Ours will take a few seconds. There's no

point in asking these big questions if you have to wait days for the answers."

The new event trends analytics software, which she is developing with Allison Rozet, a Ph.D. candidate in data science, will be tested using real-world datasets and applications supplied by a health care center and a financial transaction processing company.

"In the health care field, this could save lives," Rundensteiner said. "We could detect patterns that show how infection is spreading. We could see when, for example, staff are not putting on surgical gowns or washing their hands. We can thus see problems as they unfold, so we can see where the problems are originating. We're making better tools to get the answers we need from a growing flood of incoming information."

Provided by Worcester Polytechnic Institute