Research team formalizes novel data stream processing concept

16 November 2021, by Rachel McDowell

Watermarks, considered the most efficient mechanism for tracking how complete streaming data processing is, allow new tasks to be processed immediately after prior tasks are completed. Credit: Nathan Armistead, ORNL

A team of collaborators from the U.S. Department of Energy's Oak Ridge National Laboratory, Google Inc., Snowflake Inc. and Ververica GmbH has tested a computing concept that could help speed up real-time processing of data that stream on mobile and other electronic devices.

The concept explores the function of watermarks, considered the most efficient mechanism for tracking how complete streaming data processing is. Watermarks allow new tasks to be processed immediately after prior tasks are completed.

To better understand how watermarks might be useful, the researchers studied the computation of data streams on two different data streaming processing systems. They presented the results at the 47th International Conference on Very Large Data Bases, held in August in Copenhagen, Denmark, and virtually. The paper they presented is one of the first that formally tests and examines watermarks in a basic research setting.

"There hasn't been a clear, efficient mechanism for tracking phenomena of interest in a data stream over time and across different data processing pipelines," said Edmon Begoli, AI Systems section head in ORNL's National Security Sciences Directorate. "Watermarking is an up-and-coming concept that advances the state-of-the-art in stream processing frameworks."

Computer scientists are continually looking for ways of studying real-time data so they can better anticipate consumer needs, estimate supply and demand, and deliver more accurate information to consumers. But over the last 10 years, data management has grown increasingly challenging. This challenge is in part due to the jump in real-time computing and interactions on social media sites, in autonomous platforms like self-driving cars and on mobile devices.

To determine how different platforms might effectively process real-time data, the team compared watermarks on the two that currently enable the most advanced implementation of them: Apache Flink, an open-source stream- and batch-processing framework, and Google Cloud Dataflow, a streaming analytics service. Cloud Dataflow is a fault-tolerant platform, optimized for the parallel processing of streaming data at the global scale. Flink, on the other hand, is built for processing data streams quickly and efficiently, boasting high performance compared with Cloud Dataflow.

"We wanted to see how these perform on two different implementations and look at how they might be useful for different kinds of streaming services," Begoli said.

The researchers found that Cloud Dataflow's watermarks propagation tends to have higher latencies—delays in transferring data—and that Flink's latency grows nonlinearly as the pipeline depth and compute node count increase. However, both open-source systems, which were built by the

same community, provide a similar user experience.

Begoli said watermarks ultimately offer more flexibility than previous methods of stream processing. In the context of DOE and ORNL research, they will be useful for analyzing complex cyber events as well as collecting data from multiple sources and over various time scales, such as from sensors that measure health stats, human behaviors and movements, or environmental interactions.

"Often, there are too many complex things we want to track," Begoli said. "If you want to capture all the manifestations you're interested in and know when an event begins and ends across all sources, a concept like watermarking is very important."

In the future, the team will look at generalizing watermarks across different sources of streaming data and formalizing the performance tradeoffs emanating from different styles of implementations, such as those represented by Flink versus Cloud Dataflow architectural styles.

This research leveraged internal resources at ORNL.

More information: The paper is available as a PDF at vldb.org/pvldb/vol14/p3135-begoli.pdf

Provided by Oak Ridge National Laboratory