About the SiliconANGLE Network   Editorial Masthead

ll the mysterious Apache Flink find a sweet
t in the enterprise?

Wheatley | Feb 9, 2015 | BIG DATA, EnterpriseANGLE, Featured Articles |

che Software Foundation raised a few eyebrows last week when it
ced Apache Flink as its latest Top-Level Project (TLP). The move was
surprising because up until now very few people had even heard of the project, let alone put it into production. Flink doesn't
even have a Wikipedia page. Nevertheless, IT now ranks with popular open source data processing tools like Hadoop,
Cassandra, Lucene and Spark, suggesting it has great potential.

Don't be too disheartened if you haven't come across Apache Flink before. Flink got its start as a research project at the
Technical University of Berlin in 2009, and has only just begun making inroads outside of that continent. And even those who
do follow these kinds of things could be forgiven for not noticing it earlier – a quick look at Flink's Github page reveals it's
made little effort to distinguish itself from more prominent projects like Spark. The page unimaginatively explains Flink is a
"distributed, general-purpose data processing framework," supporting "iterations, incremental iterations and programs
consisting of large DAGs of operation".

A description on the FAQ page of the Flink site calls it, "an alternative to Hadoop's MapReduce component," that can work on
file systems other than Hadoop. That suggests that Flink is also an alternative the Spark, the enormously successful analytics
engine that some people think could be the successor to the MapReduce component of Hadoop.

The description fails to do Flink justice, for in the words of Stephan Ewen, Vice President of the Flink project and CTO of Data
Artisans GmbH, the company behind its development, it offers a lot more than rival data processing engines can muster.

Just another data engine?

"Flink, as a platform, is a new approach for unifying flexible analytics in streaming and batch data sources," Ewen said in an
interview with SD Times. "Flink's technology draws inspiration from Hadoop, MPP databases and data streaming systems, but
fuses those in a unique way [and] uses a data streaming engine to execute both batch and streaming analytics."

Just like Spark, Flink is able to digest both batch and streaming data, but it also goes a step further. It's capable of analyzing
streaming data leveraging in-memory processing to enhance its overall processing speed. Krishnan Subramanian, Founder
and Research Advisor of Rishidot Research, said lightning-fast data-crunching speed and cost-effectiveness are the biggest
weapons in Flink's arsenal.

"Flink achieves this performance advantage through some of the learning from relational databases, memory management
and query optimization," said Subramanian. "The 3Vs (volume, variety and velocity) of big data requires faster and faster
processing to meet the demands of today's business, and Apache Flink can help deliver results at the speed of today's
business."

But just how fast is Flink compared to the alternatives? Well, if you're interested in real-time analysis of real-time data, then
Flink might be just the ticket.

Flink is attempting to address one of the most relevant challenges in the data analytics pipeline: unifying the processing of
historical and real-time data, explains Leon Doitscher in Medium. That's one of the main issues with Spark, an engine for high-
speed data anlaysis. Spark can be extended to analyze data in real time, but Doitscher notes that the process is rather clunky.
That's because the technology was designed to handle batches of data rather than sequential data that's constantly flowing in.
Spark Streaming was developed to work around this problem, but the process still isn't "real, real-time," because it
necessitates accumulating data in batches for a short time before processing.

"While Spark Streaming is capable of running analytics on data streams, the actual data in its micro-batch streams are only
near real-time," said Aapo Markkanen, Principal Analyst at ABI Research. "Versus Spark, Flink's biggest promise to me seems to
about cutting down latency."

Flink is also claimed to process batch information just as efficiently as Spark, thanks to its iteration operators and a built-in
optimizer that doubles as an abstraction layer. According to Doitscher, it all adds up to an incredibly powerful data processing
engine that's capable of sifting through just about any kind of unstructured data you can throw at it.

"The combination of batch and true low-latency streaming is unique in Flink," Kostas Tzoumas, CEO of Data Artisans, told SD
Times. "Unlike pure batch or streaming engines, through its hybrid engine Flink can support both high-performance and
sophisticated batch programs as well as real-time streaming programs with low latency and complex streaming semantics."
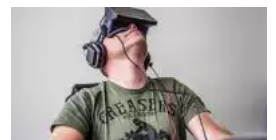
Doitscher notes that Spark users can achieve similar functionality by setting up a separate installation of Apache Storm, but that requires buying extra hardware, configuring it and running the software on top, a task made more difficult by the fact that Spark only partially supports YARN. With Flink, there is far less complexity.

Flink-forward to the future

Despite these technical advantages, Flink may struggle to gain widespread adoption given the level of support that Spark and Storm have already mustered in the enterprise. "Projects that depend on smart optimizers rarely work well in real life," said Curt Monash, an analyst at IT consultancy Monash Research, in an article in Computerworld. Monash pointed to the failure of other projects which rely on performance-enhancing tweaks, such as IBM Learning Optimizer for DB2, and Hewlett-Packard Co.'s NeoView data warehouse appliance, as a reason to be skeptical.

Nevertheless, Flink has accumulated a decent level of support in Europe. Its Github pages reveal over 75 contributors, which "is a decent amount of traction for a project that is still in the early stages," said Rishidot Research's Subramanian.

Now that its been made an Apache TPL, Flink has an even better opportunity to attract support, but doing so will be challenging. For one thing, Flink has yet to prove itself in production outside of a couple of minor European organizations, the most notable one being ResearchGate, a social network for scientists. Music streaming service Spotify Ltd. and travel software provider Amadeus IT Group, SA are also said to be testing Flink, but neither company was prepared to comment.

For Markkanen of ABI Research, Flink still needs to prove it can make a major difference in terms of performance, reliability and ease of use if it's to ever challenge the likes of Spark and Storm. But given its European roots, it seems logical that it will gain more traction there than in the U.S., at least for now.

"Flink does seem to have some early foothold in Europe, so it probably has a real chance to deliver on what it promises," Markannen said. "At the same time, Flink might well raise the bar for the alternatives. I don't think analytics will be a one-size-fits-all game. It's quite possible that Flink, Storm, Spark, etc. will all find their sweet spots."

Mike Wheatley

Mike Wheatley is a senior staff writer at SiliconANGLE. He loves to write about Big Data and the Internet of Things, and explore how these technologies are evolving and helping businesses to become more agile.

Before joining SiliconANGLE, Mike was an editor at Argophilia Travel News, an occassional contributor to The Epoch Times, and has also dabbled in SEO and social media marketing. He usually bases himself in Bangkok, Thailand, though he can often be found roaming through the jungles or chilling on a beach.

Got a news story or tip? Email Mike@SiliconANGLE.com.

About the SiliconANGLE Network   Editorial Masthead

Evolving beyond hybrid cloud to composite cloud
- Steve Chambers
Premise Hybrid cloud is being touted as the best soluti […]

MarkLogic Tries to Thread the Database Needle
- Jeff Kelly
MarkLogic held its annual customer and partner conferen […]

The Hybrid Cloud Identity Crisis
- Steve Chambers
Premise Enterprise IT has been ignited by the Digital E […]

John Cleese with a Message for Geeks
- Stuart Miniman
After thousands of interviews with technology leaders on cloud, big data and infrastructure - now for something completely different. JOHN CLEESE ON THECUBE!

theCUBE: Upcoming

RSS Error: This XML document is invalid, likely due to invalid characters. XML error: Invalid character at line 695, column 31