Where is Beam leading data processing now?

Kenneth Knowles
Apache Beam PMC Chair
kenn@apache.org

https://s.apache.org/where-is-beam-leading-now-2022
You can apply well-known functional programming optimizations to a series of MapReduces for efficiency and modularity.
Event time and processing time are orthogonal.
There is a single unified ideal of big data processing for both streaming and batch.
Any adequate big data engine can execute these universal operations.
Language-specific computation can be efficiently and generally engine-independent.
Streaming SQL is just SQL + our streaming concepts.

Schema-aware and columnar optimization belong in Beam, not just engines.
Dynamic splitting is so vastly superior to static splitting at scale, it is fundamental to big data.
Unifying splitting and work stealing with elementwise computation is a paradigm shift in expressivity.
Beam is the perfect place for engine-independent IO connectors. IO standards are the next step beyond our universal APIs.
Beam is the perfect place for AI/ML functionality to connect.
Beam is not just SDKs, runners, and connectors. Beam is part of an ecosystem with new relationships all the time.
Rich AI data engineering platform.

Orders of magnitude more connectors.

Increasingly, major ecosystem integrations outside Beam itself.

Easy "batteries included" getting started & experimentation

Python-native engines.

SDKs for new language communities.

Radical execution optimizations.
The future is bright!

Thank you!

@KennKnowles
linkedin.com/in/kennknowles
github.com/kennknowles