

Data Engineer



Spotlight

07/20/22

Build unified batch
and streaming
pipelines on popular
ML frameworks

Data Engineer



Spotlight



Sachin Agarwal

Group Product Manager,
Google Cloud

Introducing Titleist Financial

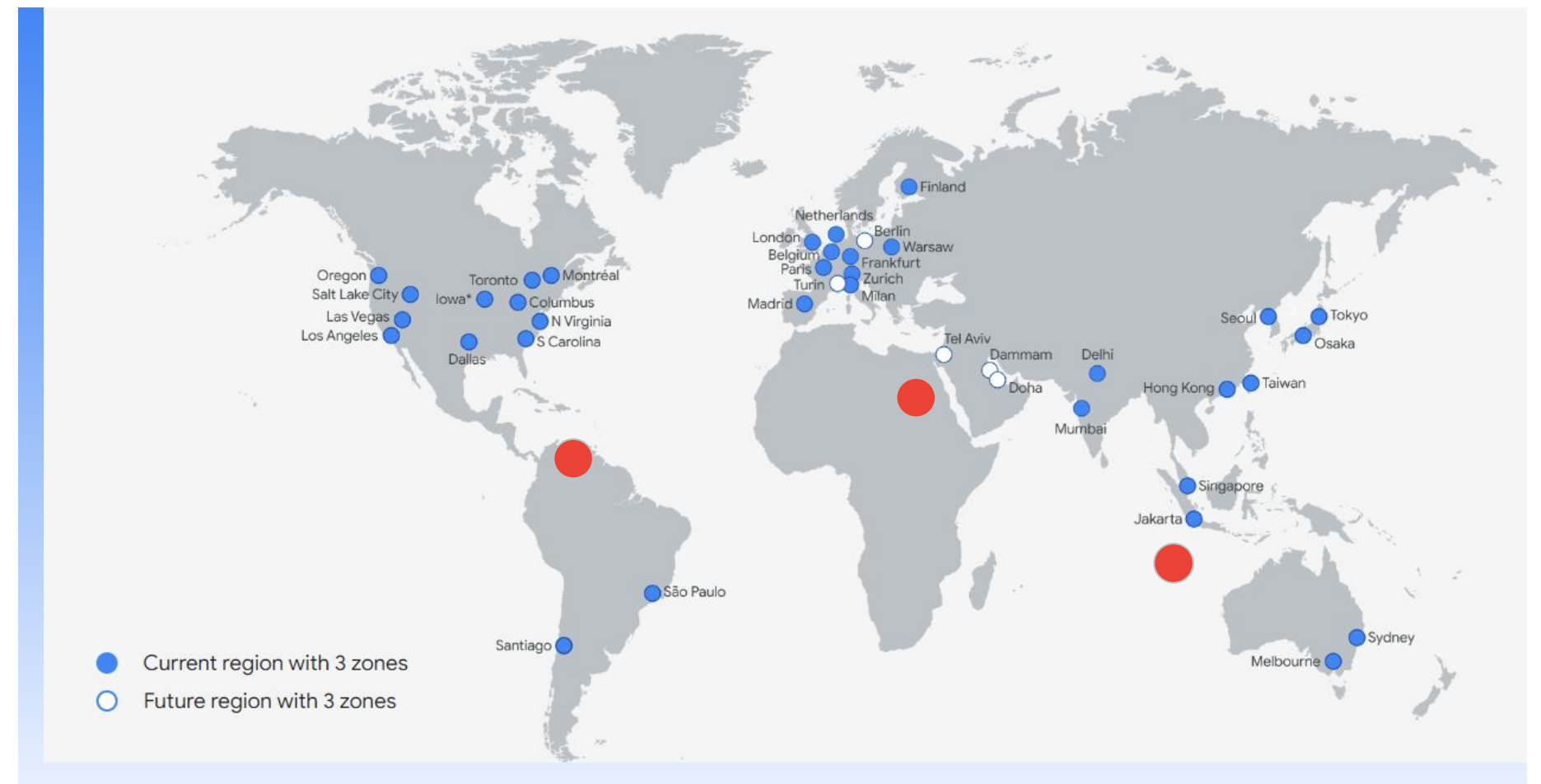
About Titleist Financial

Titleist Financial is a multinational financial services organization with teams in AMER, EMEA, and APAC. Titleist Financial has engineering teams in all locations, and various teams specialize in various areas and use different languages.

The growth engineering team based in Chicago wants to start presenting more personalized offers to customers based on real-time data. This team is new, and the engineering lead has a focus on very fast code execution with memory safety, so the growth team's preferred language is Go.

However, this new growth team needs to apply machine learning in real-time to power these personalized offers. The ML experts in the organization are based in Berlin and love Python.

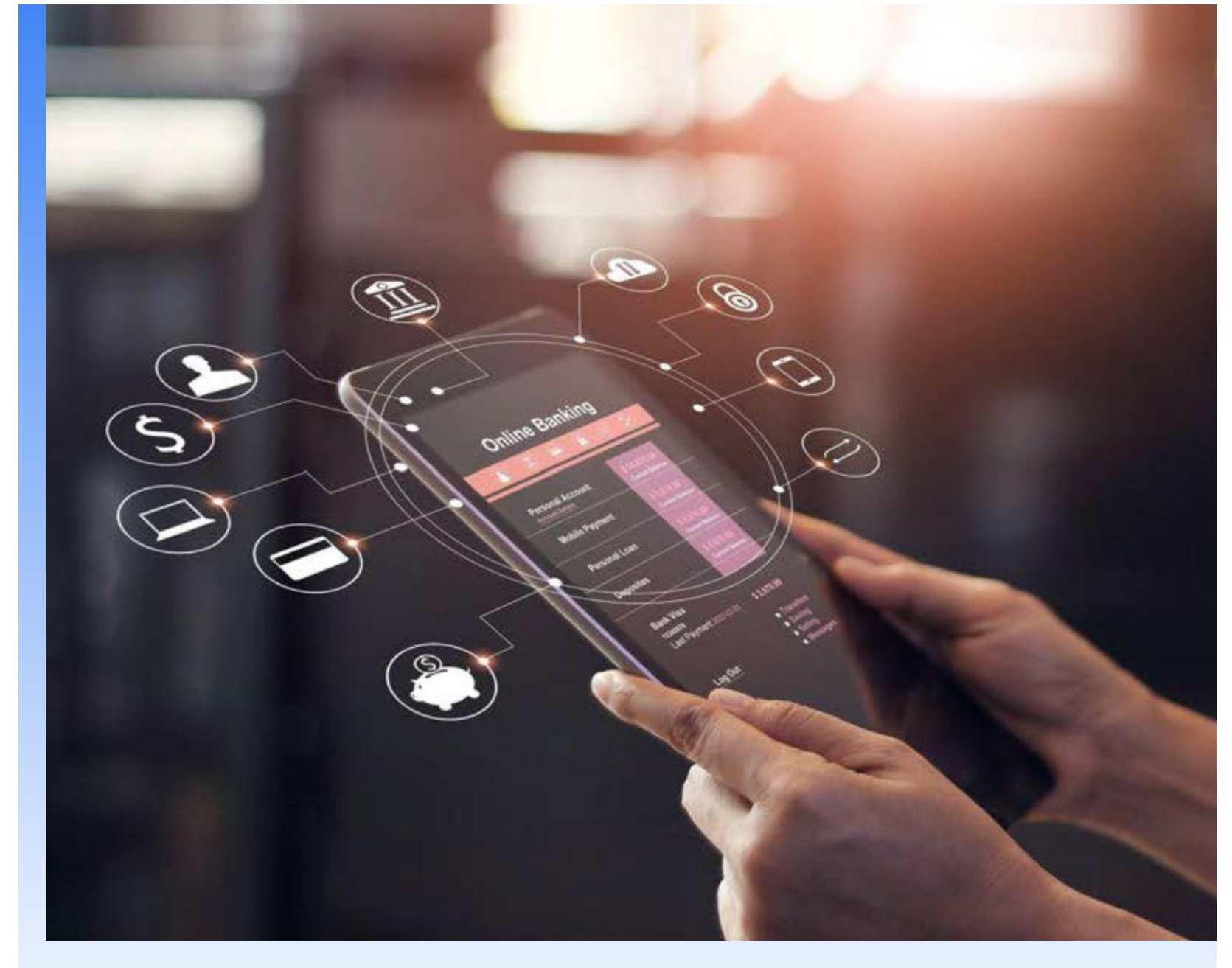
And the data experts at Titleist Financial who collect, sanitize, and verify the various sources of data that power the ML models used by Titleist are Java experts based in Hyderabad.




**One organization with
three teams in
three locations with
three different areas of expertise.**

The three team needs

- The Growth team in Chicago needs to use the same tools as the Machine Learning team in Berlin to test and deploy new experiments fast
- The Machine Learning team in Berlin needs access to all the data sources managed by the Data team in Hyderabad to power their Machine Learning models
- The Data team in Hyderabad needs the Growth team in Chicago to be responsible for the costs and infrastructure to run what they want to run





How can these distributed teams
work together seamlessly to make
sure the best offers are provided to
right customers in real-time?

Moving fast and learning fast

- The Growth team knows Go, keeps up to date with the latest Go innovations (such as generics and built-in fuzzing) and can write performant and expressive Go code - they don't want to have to give up all the benefits of that expertise
- The Go team needs to be able to use the same tools as the Python team for ML and the Java team for data access - they don't want to step down to writing lots of custom glue code, which would slow them down





Dataflow Go GA

Google Cloud Dataflow, powered by Apache Beam, is the only large scale data processing service that **natively supports the Go language**, allowing Go experts to prototype and run large-scale production data processing workloads in the language they know and love.

Dataflow Go has a number of built-in native connectors for sources and sinks such as Pub/Sub and BigQuery, but also - powered by Beam's **unique cross-language capabilities** - supports a wide range of Dataflow's Java-based I/O connectors, including Kafka and JDBC.

Dataflow Go is truly unified, like all of Apache Beam, uniquely supporting **both batch and streaming data processing with a single API**.

Individualized results in real-time

- Need to apply the output of these continuously updated Machine Learning models in real time, for each user individually
- The Berlin ML team prefers PyTorch as their ML framework of choice but is comfortable with TensorFlow, scikit-learn, and NVIDIA RAPIDS for specific use cases
- The ML team wants to use GPUs to accelerate the computation of machine learning tasks when speed matters





Dataflow ML GA

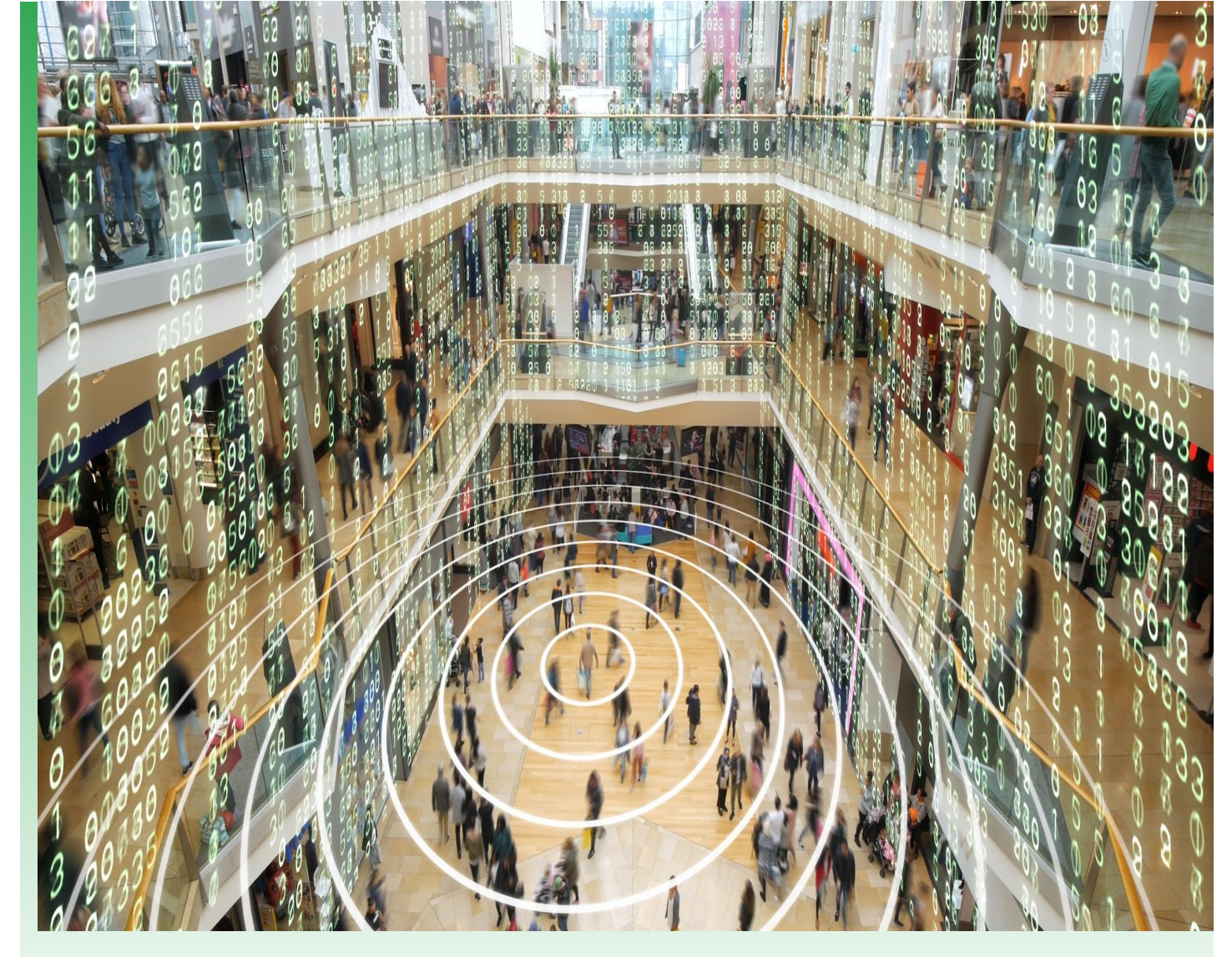
Google Cloud Dataflow uniquely allows machine learning teams **to apply the output of machine learning models** (called Inference) in real-time as part of Dataflow streaming pipelines.

Dataflow ML provides **support for PyTorch, TensorFlow, and scikit-learn out of the box**, letting machine learning teams use the framework (or frameworks) they prefer for the widest variety of tasks.

Dataflow ML also supports GPUs in addition to CPUs for machine learning tasks out of the box. Dataflow ML Inference builds on the previously released Dataframe API in the Beam Python SDK for building machine learning capabilities into Beam pipelines.

Super power, no toil

- None of the Growth team, the Machine Learning team, and the Data team has the expertise or desire to stand up, manage, and optimize infrastructure
- But the system selected has to provide low latency out of the box, and scale automatically based on changes to data or customer demand





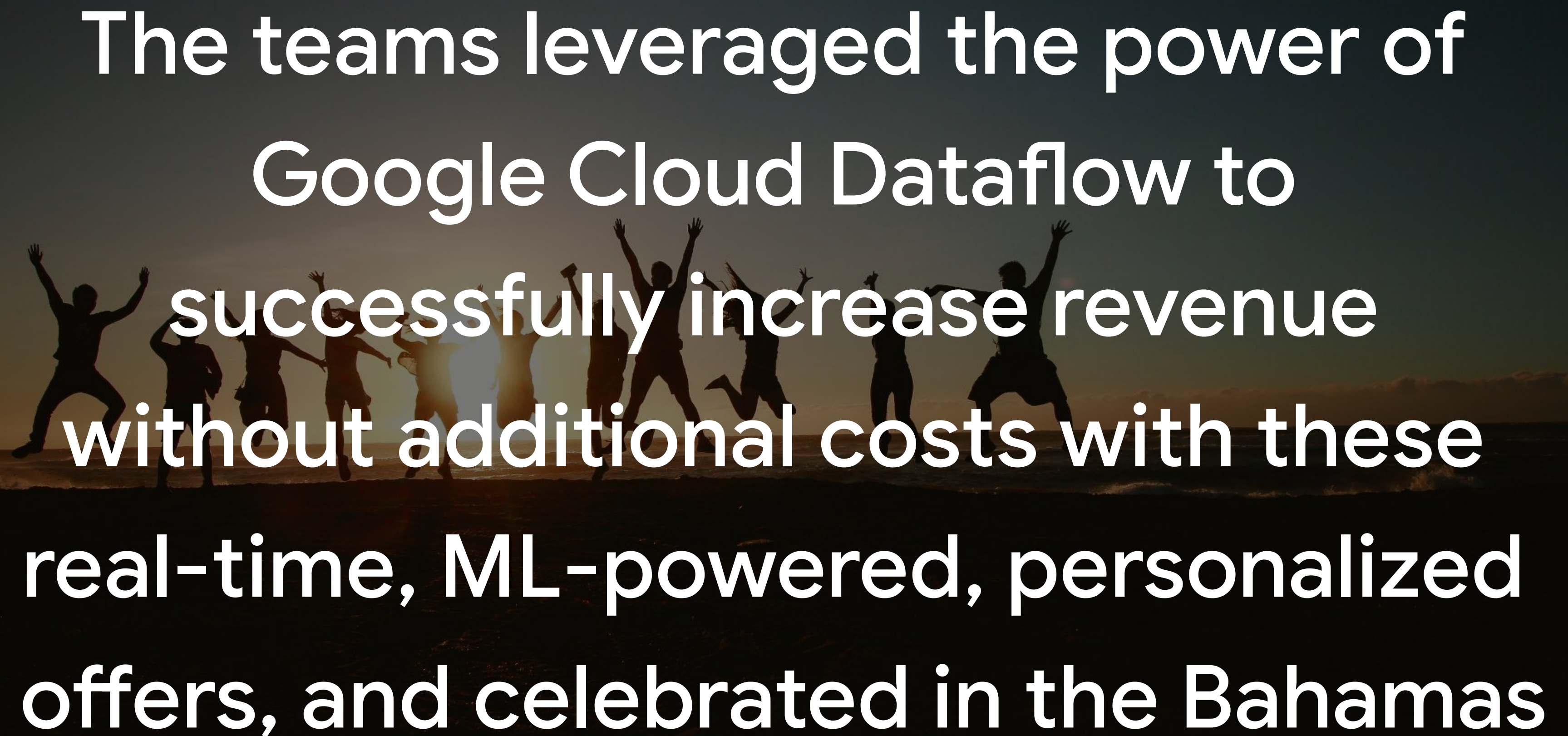
Dataflow Prime GA

Google Cloud Dataflow goes beyond just serverless to **truly no-operations for large scale data processing**.

Dataflow Prime's new right fitting applies the **best compute option for price/performance** at each and every stage of a complex data processing pipeline, using larger machines or applying GPU compute only where necessary for optimal price/performance characteristics.

Dataflow Prime's new vertical autoscaling automatically adds memory for streaming pipelines, **preventing slow or stuck streaming pipelines** before they happen.

Dataflow Prime's job visualizers and recommendations make it **easy to see and optimize pipeline behavior**.

The background of the image shows a group of people silhouetted against a bright sunset or sunrise over a beach. The people are in various celebratory poses, with their arms raised and some jumping. The sky is a mix of orange, yellow, and dark blue, and the beach is visible in the foreground.

The teams leveraged the power of
Google Cloud Dataflow to
successfully increase revenue
without additional costs with these
real-time, ML-powered, personalized
offers, and celebrated in the Bahamas

Dataflow today and tomorrow

Dataflow Go, ML, and Prime GA

Dataflow is the premier data processing solution for both batch and streaming workloads - with wide language support across languages, purpose-built machine learning functionality, and unparalleled robust price/performance, Dataflow powers businesses in all industries and geographies to meet their business needs with a truly optimized solution.

Dataflow and Beam Roadmap

Google continues to invest in the Google Cloud Dataflow service and the Apache Beam open source ecosystem. Current areas of investment include even simpler no-code solutions that build on the power of Dataflow Templates, extending Dataflow's lead in streaming capabilities, and growing the Apache Beam ecosystem.



Thank you.