White Paper

Google Streaming Analytics Platform

End-to-end, Cloud-based Streaming Analytics

By Kerry Dolan, ESG Senior IT Validation Analyst

December 2019

This ESG White Paper was commissioned by Google and is distributed under license from ESG.

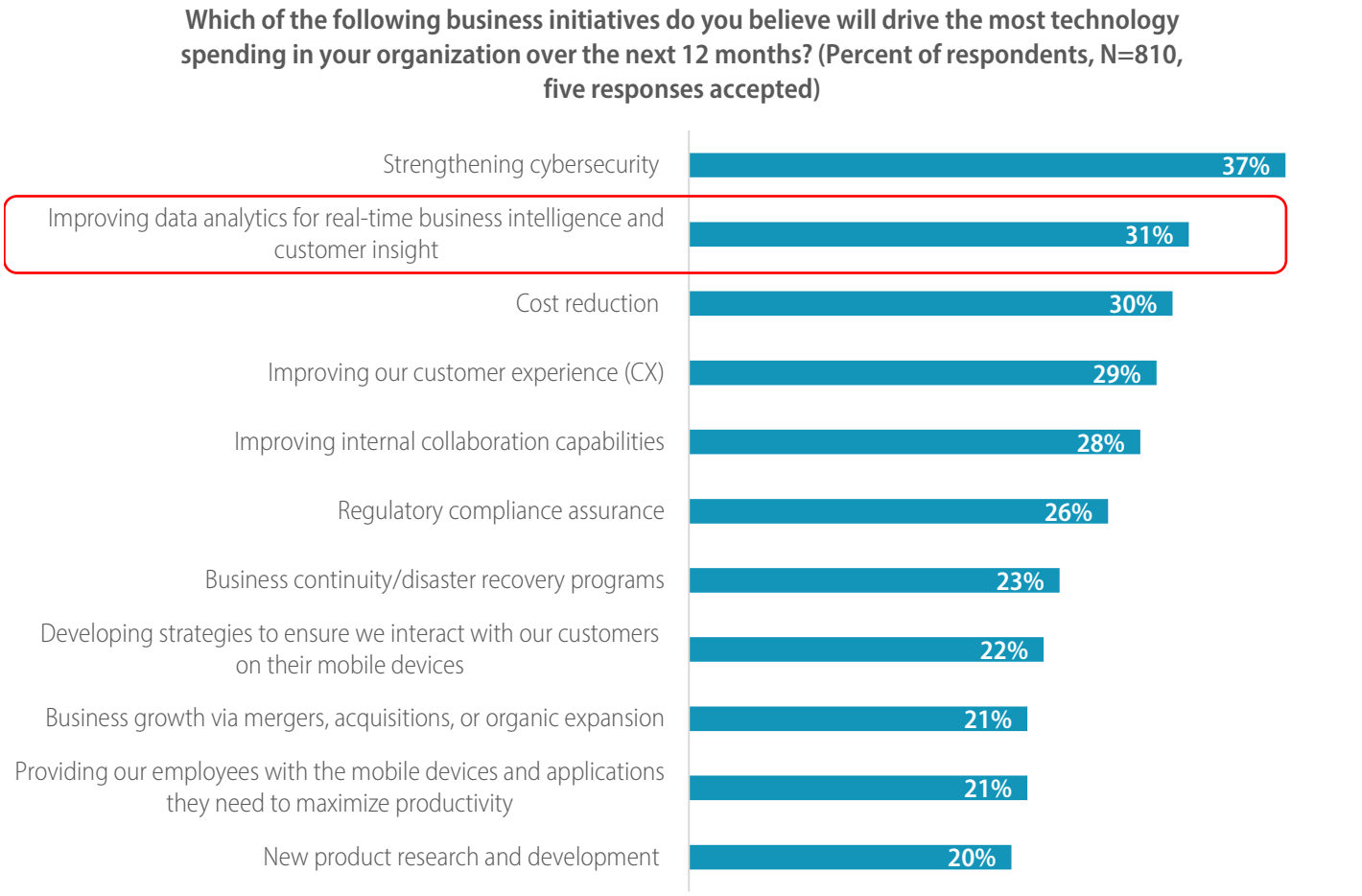
Contents

Real-time Analytics: A Business Priority

The ability to collect and use data in real time is transforming and empowering organizations in ways they never imagined. Real-time data is generated from a growing variety of sources across customer, supplier, partner, and market interactions. Messaging applications are now real-time, and sensor-enabled machines deliver a constant stream of data. Social media delivers real-time feedback and insight into consumers. Clickstream data from digital commerce can deliver predictive value. All of these data sources present the opportunity to add significant business value.

The value of mining, analyzing, and acting on this data cannot be overstated; organizations are using this data to understand customers, identify trends, design products, and head-off problems. As a result, real-time data analytics has become a key business priority. When asked what business initiatives they believed would drive the most technology spending in 2019, 31% of ESG survey respondents cited improving data analytics for real-time business intelligence and customer insight, making it the second most-cited initiative, behind strengthening cybersecurity (see Figure 1).[1]

Figure 1. Business Initiatives Driving Technology Spending



Source: Enterprise Strategy Group

Streaming analytics provides a key opportunity to analyze data in real time. To be clear, batch remains an important part of data analytics, and to the surprise of some, it's a critical component of stream analytics too, as historical and batch data can strengthen the analysis offered by real-time systems.

[1] Source: ESG Master Survey Results, *2019 Technology Spending Intentions Survey*, March 2019.

Data can be used for other strategic purposes in batch mode, *but there is a window of new opportunity if organizations can collect, process, analyze, and act on a continuing stream of data in real time*, particularly in industries like retail, financial services, media/advertising, healthcare, and utilities. The ability to collect data, instantly analyze it, and take immediate action can strengthen ties with customers and partners and enable organizations to shift more quickly in response to business conditions. Consumer and business transactions are conducted online and provide data on profiles, purchases, finances, and delivery. In retail, you can only impact behavior if you can predict in real time; if you take too long, the customer will move on. Online gaming is another example: once a user logs in, the gaming app collects a steady stream of data about play, progress made, in-game purchases, social interactions, etc. This data is used to advance the user through the experience and can be analyzed immediately by financial and marketing systems, which can then present additional real-time experiences to the gamer. Fraud detection must take place in real time for prevention. IoT data from sensors is much cheaper to analyze and maintain in real time. Event-driven data can be used for real-time responsiveness, process automation, instant interactions, targeted marketing, and myriad organization-specific processes.

Infrastructure and Skills Challenges

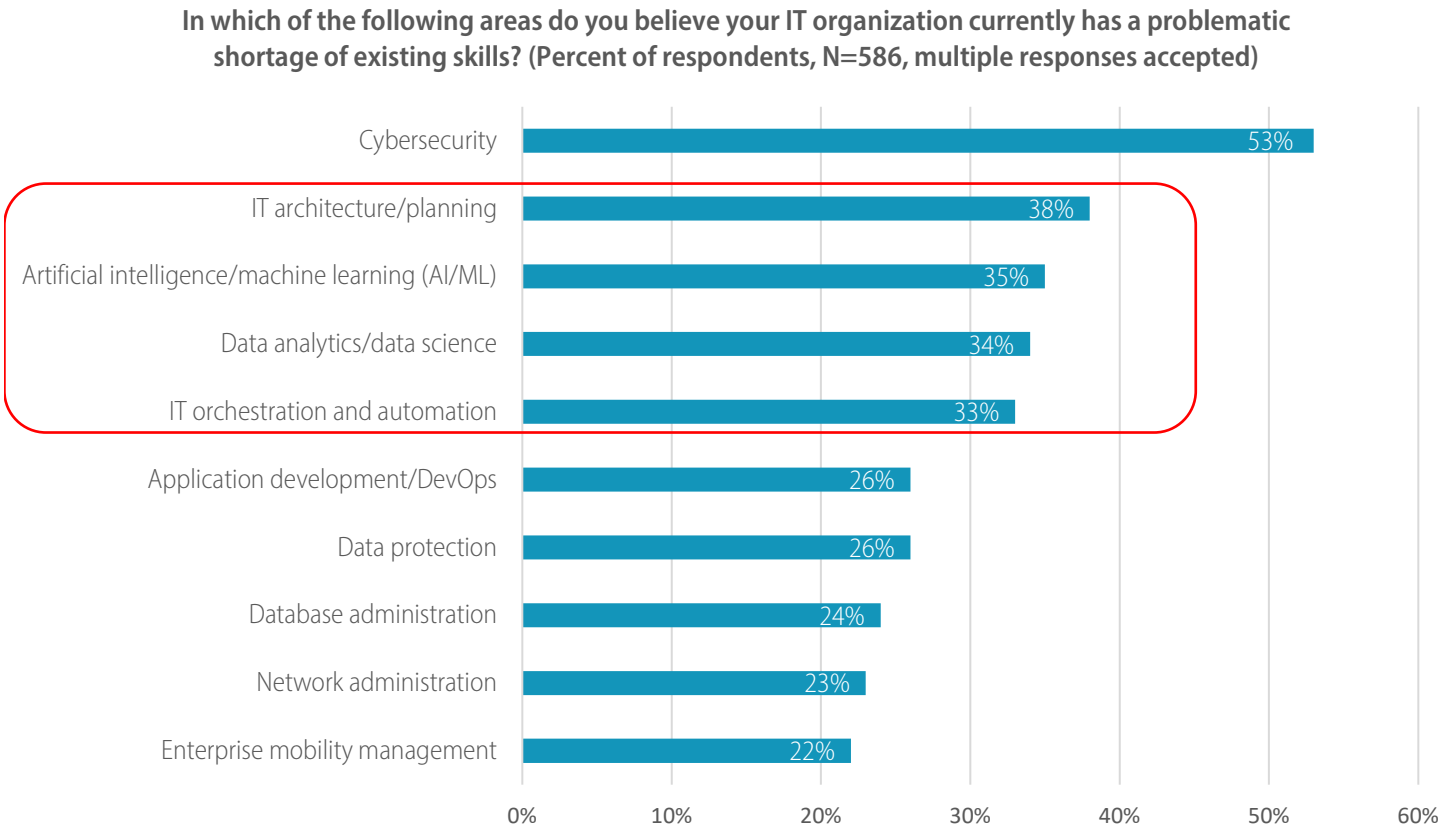
The primary challenges of streaming analytics are having the infrastructure and expertise to collect, store, and analyze the data in real time. These are similar challenges to batch processing, so what makes streaming analytics more difficult?

- It deals with very small ingest data—1 KB at a time—compared with files that will be ingested at rates of more like 20 MB at a time. Each event demands processing and delivery as soon as it happens for immediate action.
- The system needs to store and persist a continuous stream of these small events, distributing the load among clients without introducing latency.
- Jobs may run for days, months, or even years—continually rather than with a clear start and finish to the processing; with streaming data, it's hard to know when all the data has been collected. Systems must be stable and self-healing to avoid interrupted processing.
- Data must be fresh to be useful, so the analysis engine must have instant access.
- Administrators must know immediately when something is wrong so they can troubleshoot and resume operations.
- Late-arriving data causes issues when aggregate statistics need to be produced for time intervals that might not have all of the data.
- Performing complex aggregations over the time dimension is difficult.
- The arrival time of an event is not the same as the business transaction time of the event, making analytics more difficult.

It takes significant compute, storage, and networking infrastructure to deal with the heavy flow of data ingestion from a growing number of data sources, to be able to scale up and down as needed for efficiency and growth, and to process, store, and analyze it all. Plus, organizations in regulated industries must be certain that they capture every event without losing any data. Data processing engines must be flexible enough to handle different data types, and various types of application expertise are required for both infrastructure and processing/analysis applications, all using different toolsets. *Organizations want to focus on using all this data to inform decisions and actions, not on buying, building, and managing a lifecycle of applications and infrastructure for each segment of the streaming analytics process.*

Additionally, organizations are struggling to find staff with the skills they need to plan, deploy, and manage the end-to-end process. When asked in what areas they believed their IT organizations had a problematic shortage of skills, IT architecture and planning, artificial intelligence (AI)/machine learning (ML), data analytics/data science, and IT orchestration and automation were among the top five most-cited areas, surpassed only by cybersecurity (see Figure 2).[2] These contribute to why many organizations struggle to take full advantage of their data. The infrastructure costs are already daunting; when combined with the complexity of processes and applications and a lack of IT skills, many are unable to take advantage.

Figure 2. Top Ten Problematic Skills Shortages



Source: Enterprise Strategy Group

Google Streaming Analytics: An End-to-end Platform

Google offers a complete, cloud-based streaming analytics platform that provides automation, scalability, and ease of use so organizations can focus on analyzing and operationalizing their data, not on infrastructure and administration. This lets organizations quickly, easily, and cost-efficiently start using their data to drive insights. Google's services require no hardware to deploy and maintain, and no upfront costs; they include simple administration and automated scaling. Costs are limited to exactly what is needed for a job's execution, as Google's autoscaling eliminates the need to overprovision for unexpected spikes in data creation/ingestion.

The Google Streaming Analytics platform provides:

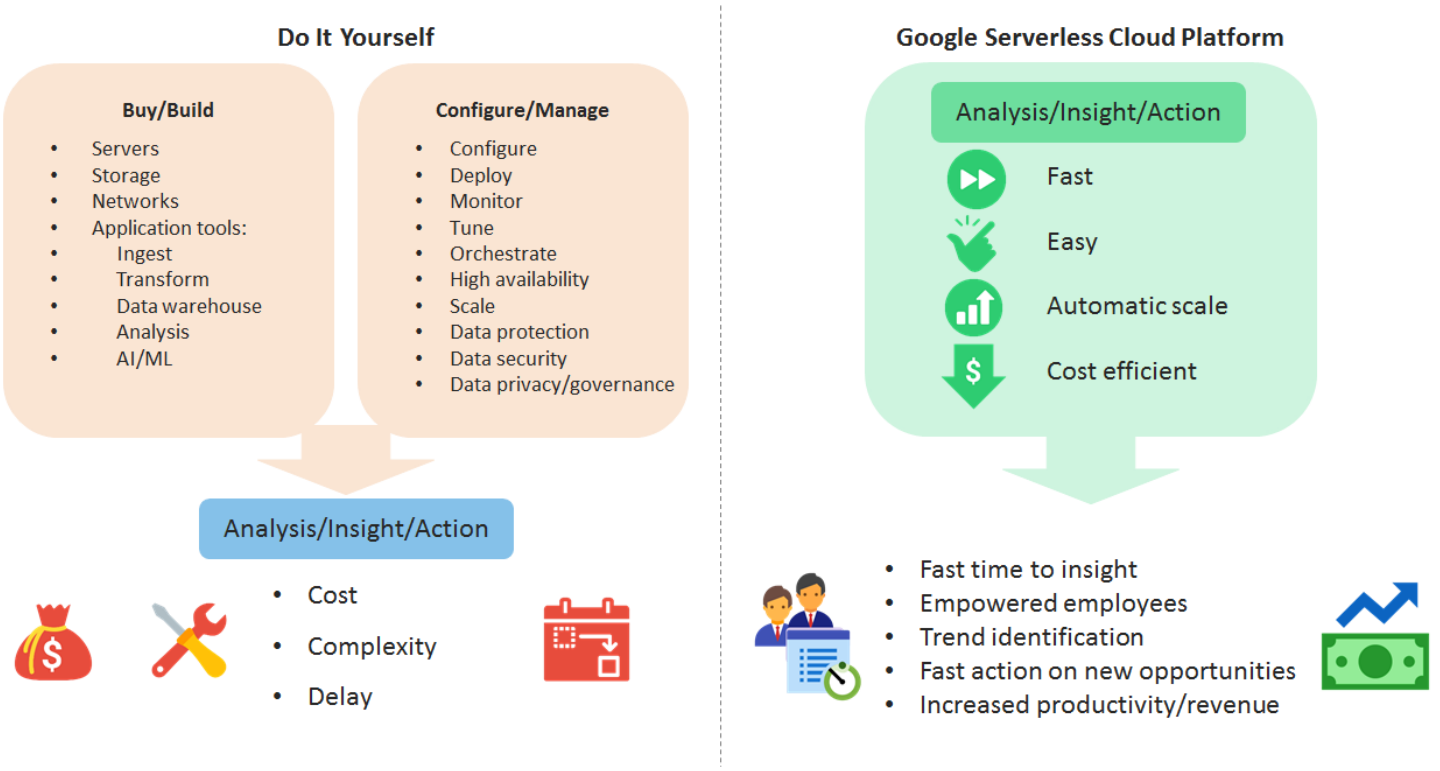
- *Robust ingestion services*. Cloud Pub/Sub takes in data and events reliably and publishes them out to multiple subscribers while reducing redundancy.

[2] ibid.

- *Unified stream and batch processing*. Cloud Dataflow changes events and data into actionable insights; there is no separate infrastructure for stream versus batch, simplifying management and reducing costs.
- *Serverless architecture*. This service-based offering can automatically scale up to handle spikes of data and back down when event volumes subside, while managing all resource-intensive provisioning and tuning tasks.
- *Comprehensive set of analysis tools*. It includes an integrated toolset across ingest, processing, and analysis versus stitching together disparate tools.
- *Flexibility for users*. Apache Beam, Dataflow's SDK, is an open source programming model that enables portability and language choice.

To make the best use of a managed service platform for streaming analytics, customers need a way to get data into the platform that's reliable, scalable, and fast. They need a processing engine that can transform that data the instant it reaches the platform. They need services on top of that data, such as SQL semantics, AI/ML, or custom application logic. They need data to be protected and secure. In regulated industries, they need controls for privacy and data governance, and they need the ability to create a duplicate of untransformed data as a full-fidelity record. The Google platform takes care of all these concerns so customers can concentrate on analysis and insights. In contrast, traditional platforms require organizations to handle configuring and deploying infrastructure; monitoring; tuning; and ensuring reliability, protection, security, and provisioning for scale—and they need to revisit these constantly throughout the lifecycle (see Figure 3).

Figure 3. DIY versus Google Serverless Streaming Analytics

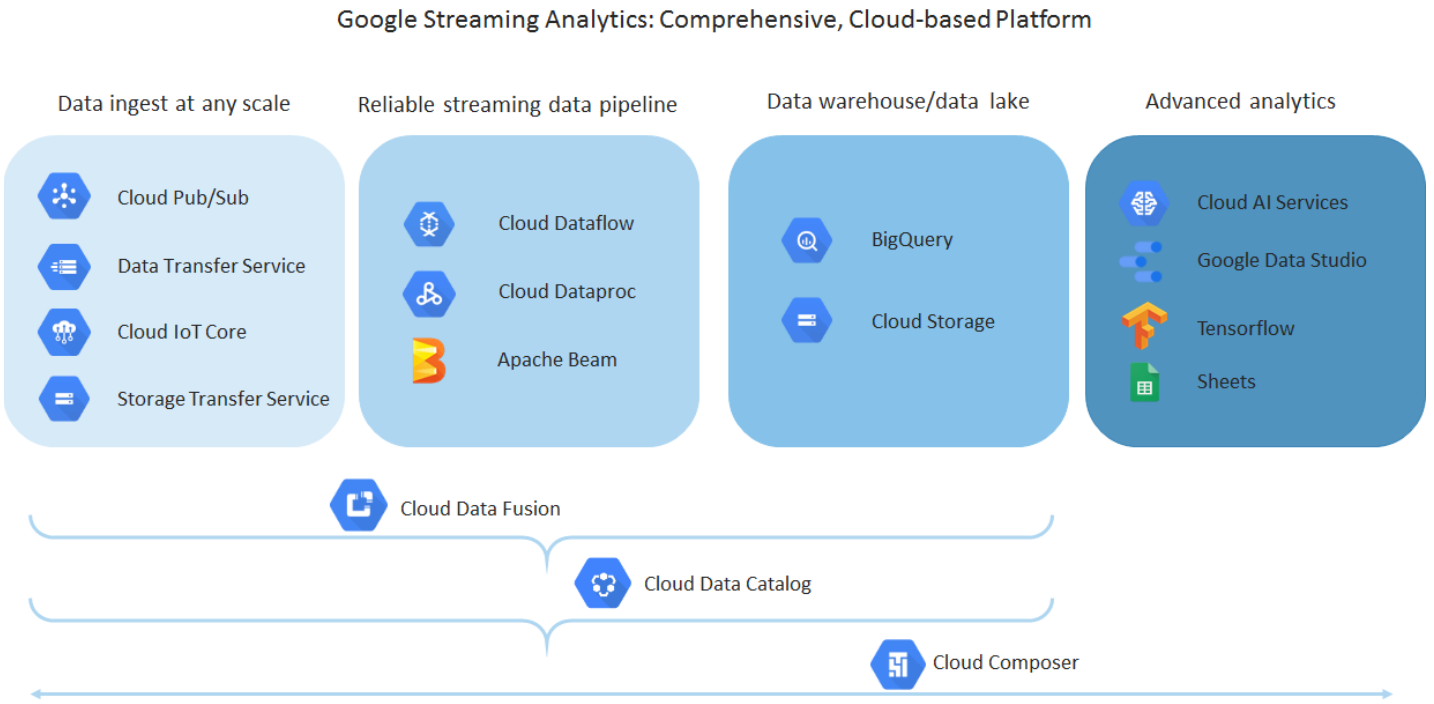


Source: Enterprise Strategy Group

Figure 4 shows the products that make up Google's end-to-end streaming analytics platform, from ingest at any scale through analysis. All parts of this platform are delivered as fully managed and integrated services, relieving

customers of the burdensome infrastructure tasks required. These services are available in every Google region around the globe.

Figure 4. Google Cloud Analytics: Comprehensive, Cloud-based Platform

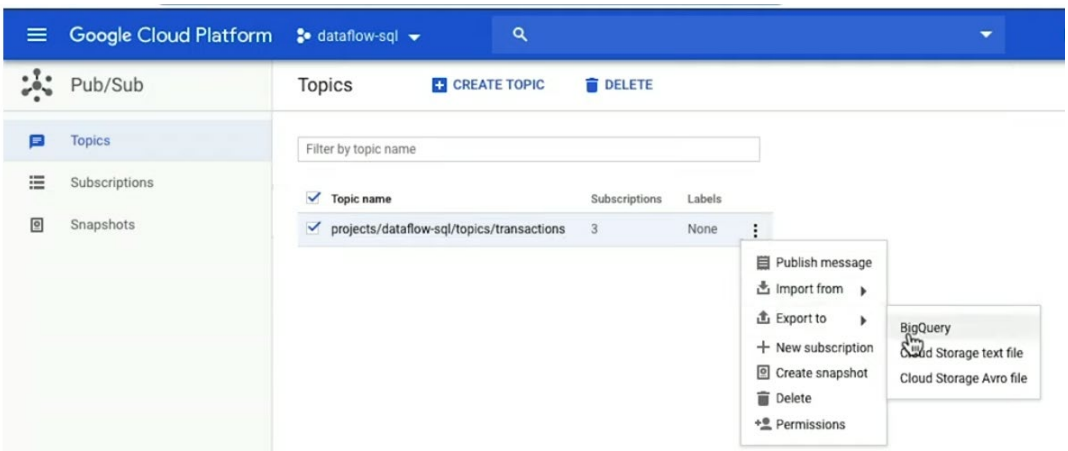


Source: Enterprise Strategy Group

Ingest

The entry point for stream analytics pipelines, ***Cloud Pub/Sub***, takes in event and data streams from media, messaging, sensors, IoT, etc., organizes them into topics, and makes them available to subscribers by topic. Organizations can publish and subscribe to events in any geography. Data is replicated synchronously across zones for up to seven days for availability. Cloud IoT Core provides device connection and management for IoT use cases.

For real-time enrichment of streaming data with broader legacy/batch data sets, Data Transfer Service and Storage Transfer Service enable organizations to ingest data from on-premises locations, SaaS applications, IoT devices, and between clouds.



Cloud Pub/Sub benefits include:

- The topic-based system enables a one-to-many model of data publication so that multiple pipelines can be built from a single stream. For example, forecasting, inventory, staffing, and billing pipelines can all subscribe

to the latest sales data to use for specific analyses. Real-time and batch processing can be mixed, reducing the numbers of patterns and pipelines to deploy, and simplifying architecture.

- The system scales automatically and rebalances workloads to support the data without additional provisioning. This is a huge benefit in terms of time savings, agility, and cost. Compute and storage can scale at different rates. In addition, customers set throughput quotas for publishing, pulling/pushing data, and administrative operations (e.g., Get, List, Create, Delete, etc.). Topics can expand globally.
- Pub/Sub supports containerization and microservices for multiple inputs and data streams within an application.
- Enterprise security is built in, including end-to-end encryption, identity and access management (IAM), and audit logging.
- Native client libraries provide data engineers with a choice of languages in which to work, and an open source API supports cross-cloud and hybrid deployments.
- Simple pricing lets customers pay only for what they consume. There is no need to estimate or monitor usage, or to pay for overhead that you might not use.

Transform/Process

Cloud Dataflow is a unified stream and batch processing engine that leverages Apache Beam as its SDK, enabling organizations to build processing pipelines with the languages they choose. Apache Beam also offers freedom from lock-in, as code written to Beam can be executed on Dataflow, Apache Spark, Apache Flink, and other "runners."

Customer Success

"Google Cloud enables us to operate at scale with ease. . . We can focus on creating new things rather than on maintaining systems and worrying about things like Black Friday [scalability]. Instead we focus on our vision, where every customer has a personal experience with the brands they love." - Qubit, customer experience/personalization company

Using the same code for batch and stream reduces both costs and complexity; the latter is especially important in light of the skills shortages in data analytics, AI/ML, and infrastructure mentioned previously. Dataflow ensures "exactly once" processing (eliminating both duplication and missed inputs) with fault-tolerant execution.

The Dataflow service includes and calls upon both compute and storage hardware, but customers don't need to know anything about them, other than that they are decoupled to save the customer money. Dataflow manages resources and schedules, scales and rebalances workloads, monitors, self-heals, and collects logs. Customers simply know their data is going to the right places for processing, and Google handles the rest. Benefits include:

- Dataflow automatically manages performance, scale, availability, security, and compliance. It offers both administrative and cost efficiency, supporting parallel data processing and charging customers only for what they consume.
- Dataflow tracks small data bits and assigns them to processing nodes with a focus on workflow scheduling and dynamic rebalancing. This is particularly helpful for stream data, which often comes in peaks and valleys. Dataflow will automatically add new workers and assign data to them if that will improve execution time, and spin down workers when the volume subsides. As shards take longer to process and begin to affect worker execution time, Dataflow will reallocate load from the struggling workers to others.

- Flexibility is built in. With Apache Beam as the SDK, Dataflow job code can be deployed in other clouds, on-premises, or using Apache Spark, Apache Flink, or other runtimes. This flexibility extends to languages as well, giving engineers choices including Java, Python, and Go.

- Dataflow and Beam also support SQL. Data analysts who are not data engineers often use SQL for streaming pipelines, so this support eliminates or reduces their reliance on data engineers to build and adjust pipelines.

- Dataflow's batch and streaming flexibility extends beyond the ability to execute both paradigms. For batch jobs that can be run overnight, Dataflow's flexible resource scheduling enables these jobs to be done at a lower cost with guaranteed start windows. This gives organizations the flexibility to move processes between stream, batch, and overnight batch as needed while optimizing for cost. For example, "first draft" route and delivery scheduling for a logistics company may occur overnight for the following business day, after which real-time inputs on delays, traffic, and package priority will reroute drivers in real time. This enables flexibility and cost savings.

Google Streaming Solution Accessibility Demonstration

As mentioned earlier, infrastructure and skills are two of the largest barriers to adopting stream analytics. We've explored how Google's autoscaling capabilities address the infrastructure barrier, but Google has also addressed the streaming skills gap through product development. Google's Dataflow SQL gives data analysts the ability to create new streaming pipelines using SQL semantics from within BigQuery, Google's data warehouse. As a result, staff members with a wide variety of skillsets can access streaming data within Google's platform, removing bottlenecks and opening up engineering resources for other jobs.

ESG viewed a demonstration of the Dataflow SQL capabilities, which showcased Pub/Sub, Dataflow, and BigQuery. The demo featured creation of a streaming job that joins data from a streaming Pub/Sub topic with a BigQuery table, populating the output table in BigQuery for immediate analysis and dashboarding. The context for this demonstration was examining sales transactions, and specifically retrieving real-time sales transactions by time-based windows.

Customer Success

"With our new data pipeline and warehouse, we are able to personalize access to large volumes of data that were not previously there. That means new insights and correlations and, therefore, better decisions and increased revenue for our customers." - AB Tasty, personalization and A/B testing company

- In the Pub/Sub UI, we created a *Transactions* topic (to include a continual stream of sales data such as who purchased a product, where, when, the price, etc.), and with a right-click, selected ***New subscription/Cloud Storage text file***.

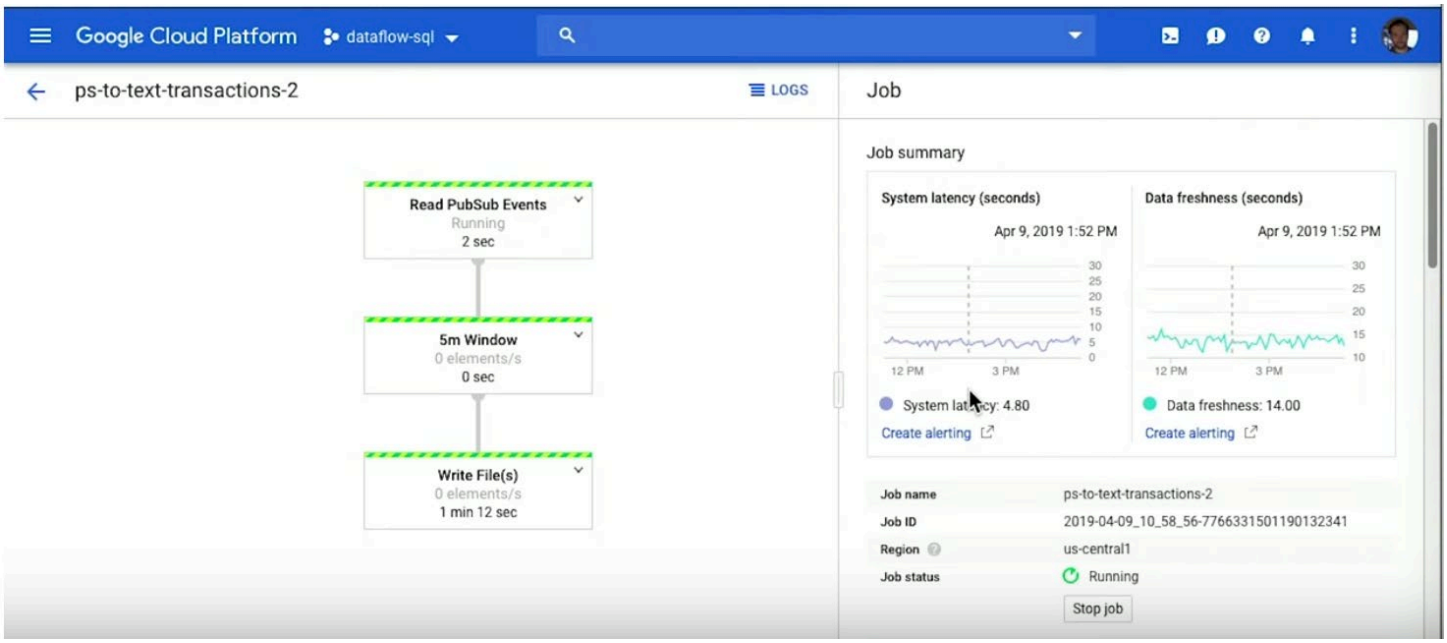
- Next, we used the Dataflow ***Create job from template*** UI to name the job and select the destination in which to store the data stream from Pub/Sub. Instructions for creating a job were viewable in the right navigation bar.

- Dataflow created a three-step job immediately: read the events, bundle them into five-minute chunks, and write them to a BigQuery table. Once in BigQuery, the data was immediately available for analysis and dashboarding.

- While the job ran, system latency and data freshness graphics were generated (see Figure 5).

- We clicked on ***Create alerting***, which took us to the Stackdriver monitoring UI, and created an alert to notify the administrator when latency exceeded 20 seconds.

Figure 5. Pub/Sub and Dataflow: Create Job with Latency and Data Freshness Graphs



Source: Enterprise Strategy Group

Data Warehouse/Analyze

Google BigQuery is a cloud-native, fully managed enterprise data warehouse supporting large-scale analytics and is a common target for streaming pipelines. It offers high-performance analysis of large data sets, with automatic scaling up and down to maximize query performance and cost. It eliminates the overhead and complexity of maintaining on-premises hardware and administration. As a cloud-native data warehouse, BigQuery also decouples compute and storage to provide cost-effective resources that are unavailable to on-prem users or to cloud data warehouses based on legacy technology.

BigQuery streaming enables transformed data to be streamed in from Dataflow one record at a time with immediate querying. BigQuery also offers a BI engine for fast, in-memory analysis of data stored in BigQuery with sub-second latency. This provides fast dashboards and is required for real-time systems with human interaction endpoints such as driving instructions or in-store alerts.

BigQuery also supports direct stream ingestion via the BigQuery Streaming API. This enables customers to deploy an ELT model that takes advantage of broadly-available SQL skills for analytic processing, which can help customers generate faster analysis or unburden data engineering resources. This can be used for automated processes, interactive querying of real-time data, or real-time BI dashboards.

Benefits include:

- Fast time to value. Customers can get their data warehouse environment up and running quickly and easily without expert system and database administrative skills.
- Speed. BigQuery speeds ingest, query, and export of petabyte-scale data sets for faster insight.

- Ease of use. Simple management includes an intuitive interface and automated scaling to petabyte scale, so that customers don't need to throttle data streaming into it. Queries finish efficiently and resources are then reallocated to other projects/users.
- Reliability and data security. Google handles geo-replication for always-on data availability and delivers continual uptime. Data protection, recovery, encryption, and IAM are also provided.
- Cost optimization. This includes predictable costs with flat rate or pay-as-you-go pricing. Because compute and storage are separated, storage can be offered at a lower cost, and customers can establish project/user resource quotas.

ESG Economic Value Validation

ESG validated a 52% reduction in three-year TCO, including cost reduction and economic benefits, for migrating an enterprise data warehouse to BigQuery versus on-premises infrastructure.[3]

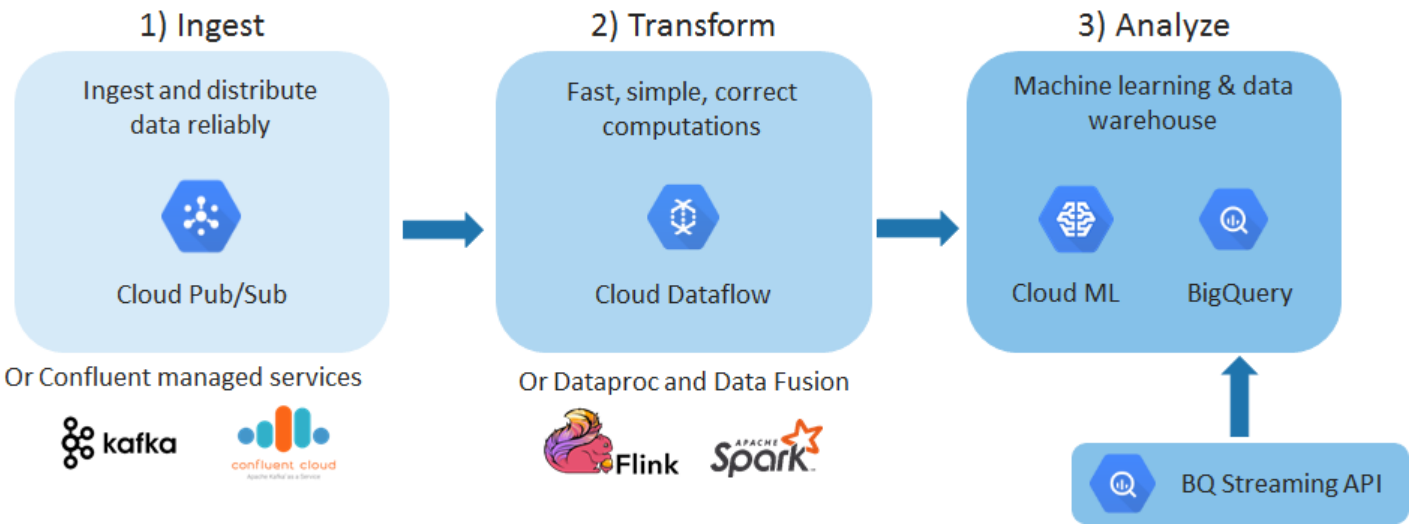
Google Alternatives to Its Prescribed Streaming Pattern

Based on the company's extensive experience running search and advertising businesses with real-time inputs, Google believes the best architecture for streaming is one that can do as much to automate infrastructure—particularly stream ingestion and stream processing—as is technically possible. ESG's survey respondents seem to agree, having listed various components of infrastructure management as areas of problematic skills shortages at their organization (see Figure 2).

However, there are users and organizations for which infrastructure and skills are not an issue. These companies may be interested in the configurability that Apache Spark provides. They may have existing Apache Kafka streaming solutions on-premises that they're looking to extend to the cloud. They may have an organizational philosophy that locks them to open source technology. Or they may want a GUI provided by a data integration tool through which they compose their streaming pipelines.

Google's broader compilation of streaming-enabled products gives customers the flexibility to mix and match technologies to achieve the optimal combination.

Figure 6. Customer Choice for Stream Data Analytics



Source: Enterprise Strategy Group

[3] For details, including information on customer successes, please see the ESG Economic Value Validation, *The Economic Advantages of Migrating Enterprise Data Warehouse Workloads to Google BigQuery*.

Cloud Dataproc

Cloud Dataproc provides cost-effective, managed processing for Hadoop and Spark environments, enabling customers to retain their familiar on-premises architecture and tools while using Google cloud storage. This high-performance, cost-effective service is easy to deploy and scale; clusters can be spun up and down as needed in minutes and are easily customizable for optimal resources on a per-job basis. For example, Dataproc offers customizable machine types, such as compute-intensive for machine learning versus standard for ad hoc analysis; these machines can read and write from the same Google cloud storage but don't compete for resources. Clusters can be tailored to use cases. For ephemeral jobs, Dataproc creates right-sized clusters, runs the jobs, and breaks down the clusters, saving data to Stackdriver to keep a record. Long-standing clusters run continuously for jobs such as streaming analytics, and also for BI, web notebooks, and ad hoc analysis. Features include auto-scaling, workflow templates, high availability mode, stable back-end storage, and low TCO.[4]

Dataproc specifically accomplishes stream processing with Apache Spark, the popular open source framework for data processing. Given the ubiquity of Spark within enterprises, both from an architecture and skills perspective, some users may prefer to accomplish stream processing within Google Cloud through Dataproc—particularly for migrations. Customers with heavy streaming requirements may then opt to transition their workloads to Dataflow after gaining more experience with the platform.

Flexibility

With Kafka, Dataflow, Dataproc, and Data Fusion, Google Cloud Platform offers an alternative to its proprietary streaming platform through an open source, fully managed solution.

Cloud Data Fusion

In Google Cloud, streaming pipelines can be deployed directly via Apache Spark from Cloud Dataproc, or indirectly through Cloud Data Fusion, Google Cloud Platform's ETL offering. Cloud Data Fusion provides code-free, visual drag-and-drop connectors to simplify data migration and transformation from on-premises and hybrid/multi-cloud environments. Customers simply point and click through sources, sinks, and transformations without any coding.

For users with limited engineering experience and a need to develop streaming pipelines, Cloud Data Fusion is a welcome tool. For example, data analysts or ETL developers can quickly build high-quality streaming pipelines to tackle their use cases without ever having to write a single line of code. Data Fusion also has the ability to publish and call upon private libraries of transformation code, meaning that even difficult tasks that require data engineers can be written once and called upon countless times by analysts.

Advanced Analytics

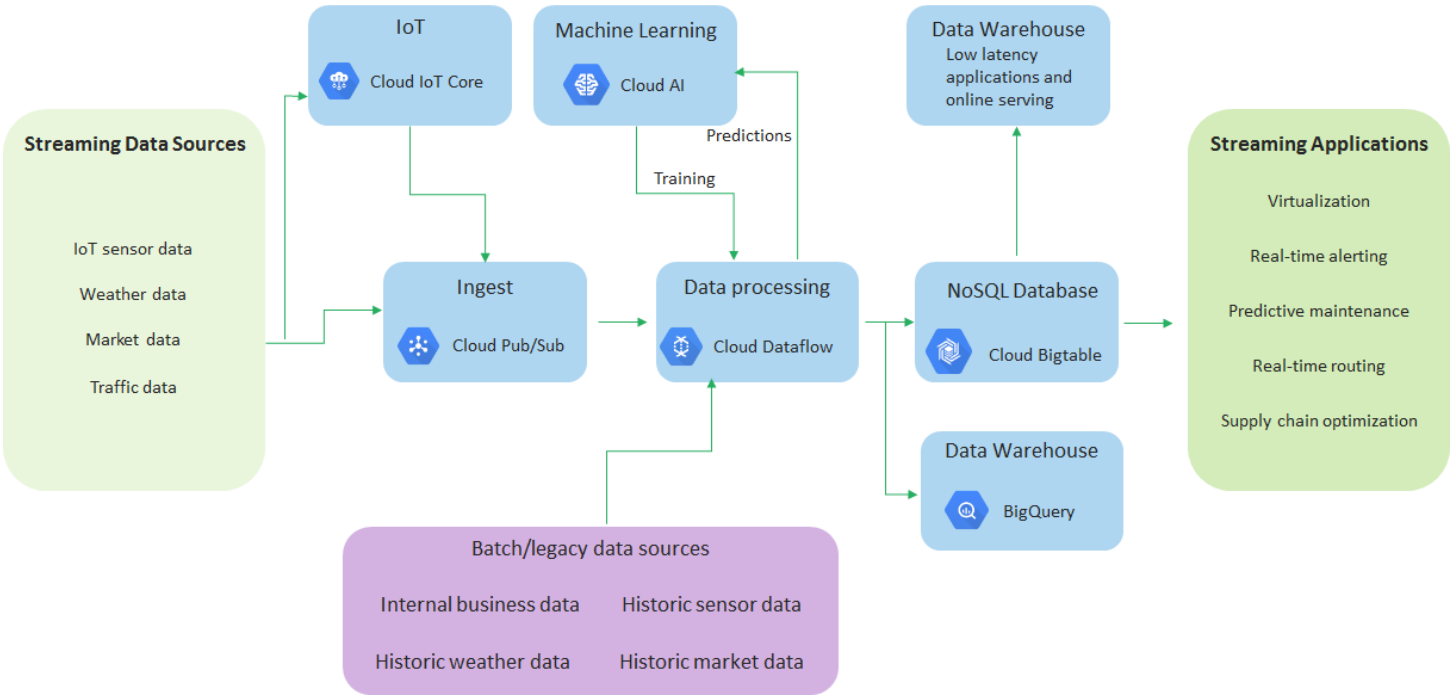
Real-time data must be analyzed in real time and immediately acted upon; otherwise, batch jobs would suffice. Google tools make it easy for users of any skill level to predict, perform, and take action on analysis in real time.

- Google offers APIs to drive processed data to analysis and action plans.
- *Cloud AutoML* simplifies training of custom machine learning models. It delivers data to automated models that have been trained without code for business-specific purposes.

[4] For additional Cloud Dataproc TCO information, see the ESG Economic Value Audit, *Analyzing the Economic Benefits of Google Cloud Dataproc Cloud-native Hadoop and Spark Platform*.

- *Cloud Machine Learning Engine* scales to meet the needs of ML models and data feeding them, removing a complex logistical burden.
- Google supports TensorFlow, designed for data experts, and delivers streaming data directly into ML processes to predict the best next option.

Figure 7. Google Streaming Analytics Architecture



Source: Enterprise Strategy Group

Customer Successes Prove the Power of Google Streaming Analytics

Customers are using Google Cloud Streaming Analytics for many tasks, from personalizing online experiences, through predicting manufacturing maintenance needs, to ensuring adequate inventory for point-of-sale transactions. The following is an overview of a customer deployment.

ITV: A Leap of Faith into a Complete Platform

ITV, the largest commercial broadcast company in the United Kingdom, provides broadcast channels and online streaming of content to millions of customers. The company captures, processes, and uses data from viewers through its apps and television programming.

In 2018, two large events loomed—the Football World cup, watched on broadcast channels and online, and the return of a hit television show. These required that ITV's data collection, processing, and analysis be able to scale without slowing down, and without increasing staff. The company needed to reliably collect 10x more data without having to do data cleaning and be ready for new applications to leverage the data. Internal customers needed analytics to power real-time product features and operational reporting, so changing the platform was a risk. The system had to be at least as accurate as the previous method, with no ingest errors, and deliver high data quality, fast processing, and room to grow.

While ITV looked at some open source solutions, they were concerned about production scale. The company initially launched a single Dataflow job, and from that success decided to fully embrace the Google Streaming Analytics platform. A high-level summary of ITV's Google stream analytics jobs includes:

- Ingest using Pub/Sub to collect data from apps and sites and scale automatically. The video playback events must deal with surges of viewers watching live events, without the high cost of pre-provisioning excess capacity. Pub/Sub topics provide the ability to have multiple consumers for a single topic, minimizing capacity and cost. Autoscaling is key.
- Dataflow job writing to BigQuery to store raw data for archive, audit, debugging, or replay of historical data. File storage was an option with a similar cost, but ITV chose BigQuery for the benefit of keeping data online and query-able.
- Dataflow jobs for user-based video playback and site interaction data are summarized and sessionized, aggregating data. Cloud Function applies formatting for hand-off to third-party organizations, providing flexibility to use stream, batch, or push mode.
- BigQuery logging jobs to store all video summaries and sessions in table format.
- Dataflow job for real-time QoS, so operations team can quickly jump on problems such as video re-buffering.
- Real-time dashboards that show data input rates by Pub/Sub topic, storage consumed, delays, etc.
- A system to deploy new code without stopping 24x7 streaming jobs. The company commented, *"GCP and especially Pub/Sub and BigQuery make it very easy to deploy new applications that use data without any risk to existing ones."*

Results

Using the Google Streaming Analytics platform, ITV now provides high-quality data and validates it, so the data is trusted by the people using it. It is available immediately to support real-time personalization. In the summer of 2018, ITV had more online viewers and successfully collected more data than ever before, while maintaining latency and reliability SLOs, with no re-engineering; the company is confident that with this platform, it can easily support the next 10x increase in data collected. In the future, the company plans to include more personalization across apps, add technology to deliver more value to advertisers, and provide a platform for data scientists to research customer behavior to drive better engagement and marketing.

> "What started as an experiment with one dataflow job has turned into a mature data platform that supports analytics, operations, and front-end product features, and sets us up for a future of more data-based applications. The leap of faith was worth it." – head of behavioral data engineering, ITV

> According to ITV's director of direct-to-consumer technology and operations, the analytics tool they built with Google Cloud Platform was created and implemented within three months, while at other major broadcasters, switching analytics tools has taken more than a year.

A few additional customer success data points demonstrate the value of Google Streaming Analytics:

- *Qubit*, a customer experience and website personalization company, tracks more than 120K events per second, including 6K tweets and 40K Google searches. The company provides more than 55 billion personalizations per month—that's eight for every person on the planet. They use Dataflow, Dataproc, and BigQuery, which stores more than five petabytes of data with 70K tables.

- When their user base tripled, *Spotify* had to make a change to handle the scale of streaming events. They needed all events (such as users creating or subscribing to playlists, listening to a song, finishing a song, etc.) delivered to a Hadoop cluster, keeping the same API, and a reliable, persistent queue for continual analysis. They launched Pub/Sub as a managed service to offload operational responsibility, and for its reliability, data retention, global availability, and REST API. Deployment was simple, and the company scaled to 700K events per second easily. Their success led them to try Dataflow to eliminate late data handling; this resulted in a 93% latency reduction. Today they are driving more than 1 TB of real-time events per day. Spotify cited clear benefits from Pub/Sub and Dataflow: the hosted service; tight integration with the complete Google ecosystem; and simple, unified batch and streaming, with the ability to write an application once and run in either mode.

- *AB Tasty*, a platform for marketing personalization on web and mobile apps, transformed from its monolithic architecture. The old architecture used five clients to process 200 events per second, and 20 GB of data per month. In 2018, they embraced the Google Streaming Analytics platform, including Pub/Sub, Dataflow, and BigQuery, and they now have 750 clients processing 600K events per second and 200TB per day. Said AB Tasty regarding savings with the Google platform, "Our legacy system would have been 12.5X more expensive to scale to today's volumes."

- *Swiss Steel*, a 175-year-old AEC company, uses Google Pub/Sub, Dataflow, and BigQuery to connect production systems and analyze sensor data. Results include improved safety with data collected from smart devices, a more efficient supply chain, and lower costs. With the Google platform, the company predicts it will save more than €500K in the next year while also gaining real-time capabilities.

- *Scotiabank*, a 186-year old bank in 50 countries with 23M customers, used GCP to create a log processing solution and a serverless, secure, auto-scaling event stream that increased scale by more than 10x. The company commented, "The transformation to standardized data- and event-driven operations is foundational to our strategy for transforming the bank. Pub/Sub and Dataflow are foundational to the implementation of that strategy. We've been really happy with the performance, the cost, the reliability of the systems."

The Bigger Truth

Organizations are generating and collecting real-time data at massive scale, from retail points of sale, factory sensors, gaming applications, web transactions, and more. There are insights buried in every event stream; that data can change an organization's trajectory—if they can only get to it quickly, make it usable, and analyze it. There are also opportunities to interact in real time while customers are in an app, through website personalization, advertising based on purchase/clickstream history, gaming interactions, etc., that provide additional revenue sources and augment existing ones.

That's the value of streaming analytics. It's about grabbing data in motion, getting immediate insight, and taking immediate action. Organizations that can respond immediately to these events as they happen have a

competitive advantage. But many organizations don't have the infrastructure, expertise, or resources to do it themselves; they remain on the sidelines, missing valuable details that could impact business.

Setting up an infrastructure to handle that end-to-end process is difficult, time consuming, and expensive. It is complex to manage not only the infrastructure, but also the applications required for processing and analysis. The system must be able to scale up and down easily, but without the high cost of over-provisioning to support growth.

Google's streaming analytics platform provides customers with a powerful, service-based solution that gets customers up and running quickly, easily, and cost-efficiently. This enables customers to focus on the analysis, insights, and actions relevant to their businesses, not on provisioning, configuring, deploying, monitoring, tuning, scaling, and protecting their infrastructure and data. Instead of silos built for different parts of the process, the Google Streaming Analytics platform helps customers build more responsive businesses, modernize data warehouses, protect and govern data at scale, and turn data into actionable intelligence.

Google can help speed time-to-insight and subsequent insight-to-action time so your business can be data-driven and can automatically scale up and down to ensure cost-efficiency. Pub/Sub can ingest and Dataflow can stream process millions of events per second and load to BigQuery for analysis. Google's platform supports open source tools such as Beam, Kafka, Spark, Flink, and Hadoop so customers are not locked into a single vendor; in addition, Google Streaming Analytics can support on-premises and other cloud deployments through tools such as Cloud Data Fusion.

The goal of streaming analytics is to deliver immediate insight and action on data as it is produced and collected. If you are in search of a way to take advantage of the hidden insights in your data, Google's powerful, serverless platform can get you up and running quickly, easily, and cost-effectively.