Dataflow

Unified stream and batch data processing that's serverless, fast, and cost-effective.

New customers get $300 in free credits to spend on Dataflow.

- Fully managed data processing service
- Automated provisioning and management of processing resources
- Horizontal autoscaling of worker resources to maximize resource utilization
- OSS community-driven innovation with Apache Beam SDK
- Reliable and consistent exactly-once processing

Dataflow is a managed service for executing a wide variety of data processing patterns. The documentation on this site shows you how to deploy your batch and streaming data processing pipelines using Dataflow, including directions for using service features.

The Apache Beam SDK is an open source programming model that enables you to develop both batch and streaming pipelines. You create your pipelines with an Apache Beam program and then run them on the Dataflow service. The Apache Beam documentation provides in-depth conceptual information and reference material for the Apache Beam programming model, SDKs, and other runners.

BENEFITS

Streaming data analytics with speed

Dataflow enables fast, simplified streaming data pipeline development with lower data latency.

Simplify operations and management

Allow teams to focus on programming instead of managing server clusters as Dataflow’s serverless approach removes operational overhead from data engineering workloads.

Reduce total cost of ownership

Resource autoscaling paired with cost-optimized batch processing capabilities means Dataflow offers virtually limitless capacity to manage your seasonal and spiky workloads without overspending.

Key features

Autoscaling of resources and dynamic work rebalancing

Minimize pipeline latency, maximize resource utilization, and reduce processing cost per data record with data-aware resource autoscaling. Data inputs are partitioned automatically and constantly rebalanced to even out worker resource utilization and reduce the effect of “hot keys” on pipeline performance.

Flexible scheduling and pricing for batch processing

For processing with flexibility in job scheduling time, such as overnight jobs, flexible resource scheduling (FlexRS) offers a lower price for batch processing. These flexible jobs are placed into a queue with a guarantee that they will be retrieved for execution within a six-hour window.

Ready-to-use real-time AI patterns

Enabled through ready-to-use patterns, Dataflow’s real-time AI capabilities allow for real-time reactions with near-human intelligence to large torrents of events. Customers can build intelligent solutions ranging from predictive analytics and anomaly detection to real-time personalization and other advanced analytics use cases.

All features

Vertical autoscaling - new in Dataflow Prime	Dynamically adjusts the compute capacity allocated to each worker based on utilization. Vertical autoscaling works hand in hand with horizontal autoscaling to seamlessly scale workers to best fit the needs of the pipeline.
Right fitting - new in Dataflow Prime	Right fitting creates stage-specific pools of resources that are optimized for each stage to reduce resource wastage.
Smart diagnostics - new in Dataflow Prime	A suite of features including 1) SLO-based data pipeline management, 2) Job visualization capabilities that provide users a visual way to inspect their job graph and identify bottlenecks, 3) Automatic recommendations to identify and tune performance and availability problems.

Streaming Engine	Streaming Engine separates compute from state storage and moves parts of pipeline execution out of the worker VMs and into the Dataflow service back end, significantly improving autoscaling and data latency.
Horizontal autoscaling	Horizontal autoscaling lets the Dataflow service automatically choose the appropriate number of worker instances required to run your job. The Dataflow service may also dynamically reallocate more workers or fewer workers during runtime to account for the characteristics of your job.
Dataflow Shuffle	Service-based Dataflow Shuffle moves the shuffle operation, used for grouping and joining data, out of the worker VMs and into the Dataflow service back end for batch pipelines. Batch pipelines scale seamlessly, without any tuning required, into hundreds of terabytes.
Dataflow SQL	Dataflow SQL lets you use your SQL skills to develop streaming Dataflow pipelines right from the BigQuery web UI. You can join streaming data from Pub/Sub with files in Cloud Storage or tables in BigQuery, write results into BigQuery, and build real-time dashboards using Google Sheets or other BI tools.
Flexible Resource Scheduling (FlexRS)	Dataflow FlexRS reduces batch processing costs by using advanced scheduling techniques, the Dataflow Shuffle service, and a combination of preemptible virtual machine (VM) instances and regular VMs.

Dataflow templates	Dataflow templates allow you to easily share your pipelines with team members and across your organization or take advantage of many Google-provided templates to implement simple but useful data processing tasks. This includes Change Data Capture templates for streaming analytics use cases. With Flex Templates, you can create a template out of any Dataflow pipeline.
Notebooks integration	Iteratively build pipelines from the ground up with Vertex AI Notebooks and deploy with the Dataflow runner. Author Apache Beam pipelines step by step by inspecting pipeline graphs in a read-eval-print-loop (REPL) workflow. Available through Google's Vertex AI, Notebooks allows you to write pipelines in an intuitive environment with the latest data science and machine learning frameworks.
Real-time change data capture	Synchronize or replicate data reliably and with minimal latency across heterogeneous data sources to power streaming analytics. Extensible Dataflow templates integrate with Datastream to replicate data from Cloud Storage into BigQuery, PostgreSQL, or Cloud Spanner. Apache Beam's Debezium connector gives an open source option to ingest data changes from MySQL, PostgreSQL, SQL Server, and Db2.
Inline monitoring	Dataflow inline monitoring lets you directly access job metrics to help with troubleshooting

	batch and streaming pipelines. You can access monitoring charts at both the step and worker level visibility and set alerts for conditions such as stale data and high system latency.
Customer-managed encryption keys	You can create a batch or streaming pipeline that is protected with a customer-managed encryption key (CMEK) or access CMEK-protected data in sources and sinks.
Dataflow VPC Service Controls	Dataflow’s integration with VPC Service Controls provides additional security for your data processing environment by improving your ability to mitigate the risk of data exfiltration.
Private IPs	Turning off public IPs allows you to better secure your data processing infrastructure. By not using public IP addresses for your Dataflow workers, you also lower the number of public IP addresses you consume against your Google Cloud project quota.