# Big Data and Machine learning

Google believes that in the future, every company will be a data company. Because making the fastest and best use of data is a critical source of competitive advantage.

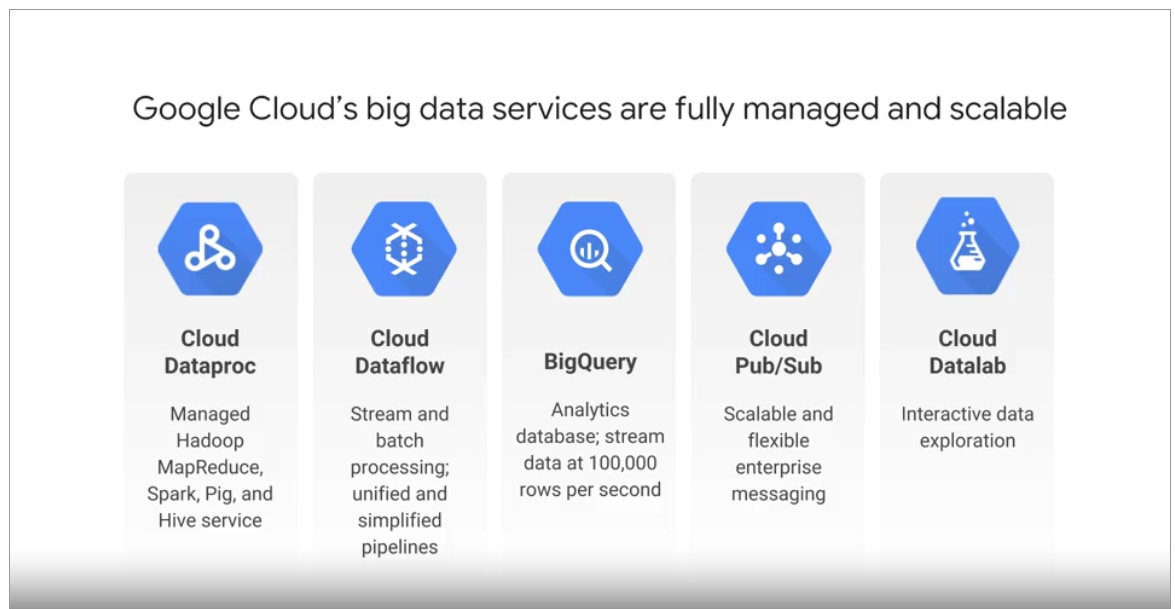
Google Cloud provides a way for everybody to take advantage of Google's investments in infrastructure and data processing innovation.

Google Cloud has automated out the complexity of building and maintaining data and analytics systems. In this module, I'll tell you about Google's technologies for getting the most out of data fastest.

Whether it's real time analytics or machine learning. These tools are intended to be simple and practical for you to embed in your applications so that you can put data into the hands of your domain experts and get insights faster.

# Google Cloud Big Data Platform

Google Cloud Big Data Solutions are designed to help you transform your business and user experiences with meaningful data insights. We like to call it an Integrated Serverless Platform. What does that mean? Serverless means you don't have to worry about provisioning Compute Instances to run your jobs.



The services are fully managed, and you pay only for the resources you consume. The platform is integrated, so that GCP data services work together to help you create custom solutions.

## Cloud Dataproc

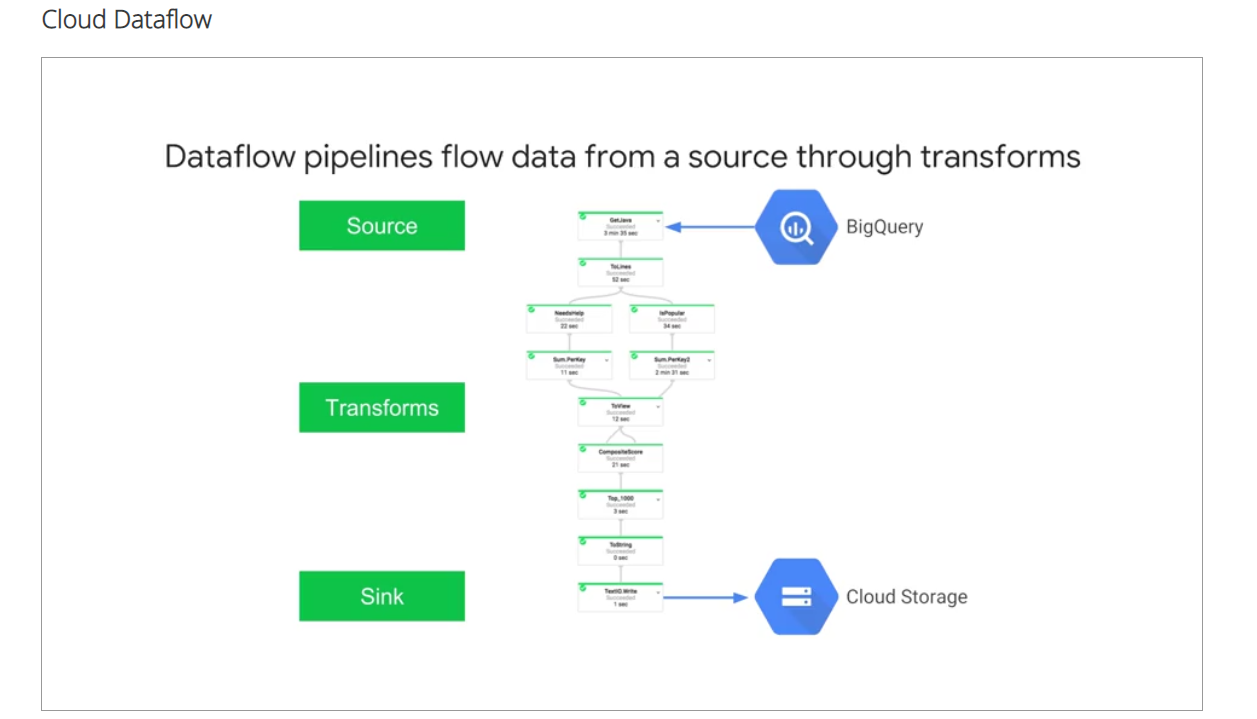
Apache Hadoop is an open source framework for big data. It is based on the MapReduce programming model which Google invented and published.

The MapReduce model is, at its simplest, means that one function, traditionally called the "Map function," runs in parallel with a massive dataset to produce intermediate results. And another function, traditionally called the "Reduce function," builds a final result set based on all those intermediate results.

The term "Hadoop" is often used informally to encompass Apache Hadoop itself, and related projects such as Apache Spark, Apache Pig, and Apache Hive. Cloud Dataproc is a fast, easy, managed way to run Hadoop, Spark, Hive, and Pig on Google Cloud Platform. All you have to do is request a Hadoop cluster. It will be built for you in 90 seconds or less, on top of Compute Engine virtual machines whose number and type you control. If you need more or less processing power while your cluster is running, you can scale it up or down. You can use the default configuration for the Hadoop software in your cluster or you can customize it. And you can monitor your cluster using Stackdriver. Running on-premises, Hadoop jobs requires a capital hardware investment. Running these jobs in Cloud Dataproc, allows you to only pay for hardware resources used during the life of the cluster you create. Although the rate for pricing is based on the hour, Cloud Dataproc is billed by the second. Our Cloud Dataproc clusters are billed in one-second clock-time increments, subject to a one minute minimum billing. So, when you're done with your cluster, you can delete it, and billing stops. This is much more agile use of resources than on-premise hardware assets. You can also save money, by telling Cloud Dataproc to use preemptible Compute Engine instances for your batch processing. You have to make sure that your jobs can be restarted cleanly, if they're terminated, and you get a significant break in the cost of the instances. At the time this video was made, preemptible instances were around 80 percent cheaper. Be aware that the costs of the Compute Engine instances isn't the only component of the cost of a Dataproc cluster, but it's a significant one. Once your data is in a cluster, you can use Spark and Spark SQL to do data mining. And you can use MLib, which is Apache Spark's machine learning libraries to discover patterns through machine learning.

# Cloud Dataflow

Cloud Dataproc is great when you have a data set of known size or when you want to manage your cluster size yourself. But what if your data shows up in real time or it's of unpredictable size or rate? That's where Cloud Dataflow is particularly a good choice. It's both a unified programming model and a managed service and it lets you develop and execute a big range of data processing patterns: extract, transform, and load batch computation and continuous computation.



You use Dataflow to build data pipelines. And the same pipelines work for both batch and streaming data. There's no need to spin up a cluster or to size instances.

Cloud Dataflow fully automates the management of whatever processing resources are required.

Cloud Dataflow frees you from operational tasks like resource management and performance optimization. In this example, Dataflow pipeline reads data from a big query table, the Source, processes it in a variety of ways, the Transforms, and writes its output to a cloud storage, the Sink. Some of those transforms you see here are map operations and some are reduce operations. You can build really expressive pipelines. Each step in the pipeline is elastically scaled. There is no need to launch and manage a cluster. Instead, the service provides all resources on demand. It has automated and optimized worked partitioning built in, which can dynamically rebalance lagging work. That reduces the need to worry about hotkeys. That is, situations where disproportionately large chunks of your input get mapped to the same cluster. People use Dataflow in a variety of use cases. As we've discussed, it's a general purpose ETL tool and its use case as a data analysis engine comes in handy in things like fraud detection and financial services, IoT analytics and manufacturing, healthcare and logistics and click stream, point of sale and segmentation analysis in retail. And because those pipelines, we saw can orchestrate multiple services even external services. It can be used in real time applications such as personalizing gaming user experiences.