10. Challenges facing modern data pipelines

Introduction

Software engineers have a central responsibility for building efficient, scalable, and secure pipelines. Data may come in from all various data sources inside modern organizations.

Message-oriented architectures with Cloud Pub/Sub

The first part of the pipeline puzzle manages vast amounts of streaming data that do not come from a single organized database. They could stream from a thousand or a million separate events, all asynchronously communicating. A common use case where you see this trend happening is for applications using IoT or the Internet of Things. IoT devices may be sensors on Go Jacks drivers' motorcycles, transmitting location data every 30 seconds for each driver, or placing temperature sensors around a data center to automate and monitor the heating and cooling costs.

Pub / Sub is an asynchronous messaging service designed for optimal reliability and scalability. The service is based on a key component of Google's infrastructure that has relied on many Google products for over a decade. Google products including Ads, Search and Gmail use this network to send more than half a billion messages per second, for a total of more than 1TB / s of data.

Serverless Application

Serverless implies no upfront provisioning, no server maintenance, and the pay-what-you-use system for application development. IT has had to manage servers for application hosting (compute) and application databases since the beginning of Client-Server development. By debuting with Google App Engine in 2008, the first truly serverless computing service, Google Cloud has always believed in the Serverless dream. Google has since developed more serverless packages in both app development and analytics.

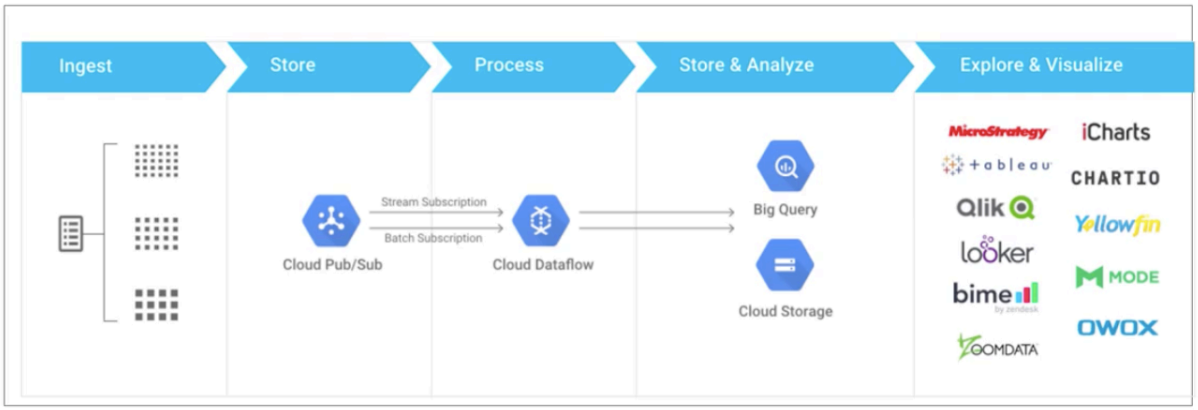

Figure 1. Google Cloud Serverless Big Data Pipeline

This guarantees the transmission of messages at least once and passes them on to subscribing applications, so no provisioning is needed. Where there is a lot or none of the messages at all, Pub / Sub must scale up to accommodate the demand.

Principle

A Pub / Sub service has some main concepts to it:

- ***Message***: data that is passing through the process.
- ***Topic***: an object identified as representing a message stream.
- ***Subscription***: an entity named that represents an interest in receiving messages on a given topic.
- ***Publisher*** (also known as a producer): produces messages and delivers (publishes) them on a given topic to the messaging service.
- ***Subscriber*** (also known as a consumer): receives messages about a given subscription.

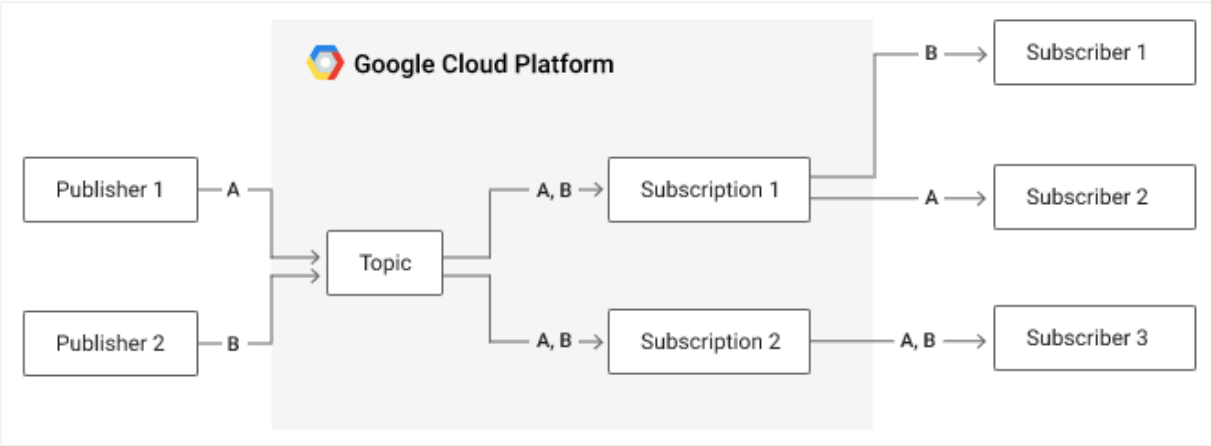

Figure 2. The basic message stream through Pub / Sub (Source; Google Cloud)

Analyzing Quality and performance

A messaging service such as Pub / Sub could be measured in three aspects based on its performance: scalability, availability, and latency. These three factors frequently contradict each other and involve compromises on one to boost the other two.

Designing streaming pipelines with Apache Beam

What is Apache Beam

Apache Beam is an open-source, centralized platform for both batch identification and the streaming of data-parallel pipelines. You create a program that describes the pipeline using one of the open-source Beam SDKs. The pipeline is then run by one of the distributed processing back-ends provided by Beam, which include Apache Flink, Apache Spark, and Google Cloud Dataflow.

Beam is particularly useful for Embarrassingly Parallel data processing projects, in which the problem can be broken down into several smaller bundles of data that can be processed independently and in parallel. You may also use Beam for tasks of Extracting, Converting, and Loading (ETL) and pure data integration (Apache.org, 2020).

You can find details on the Apache Beam web site on:

1. **How to design your pipeline**: Shows how to identify the layout of your pipeline, how to select which transforms to apply to your data, and how to establish your methods for input and output.
1. **How to build your pipeline**: illustrates the process of using the classes in the Beam SDKs and the steps needed to develop a pipeline.
2. **How to test the pipeline**: Offers the best pipeline testing practices.

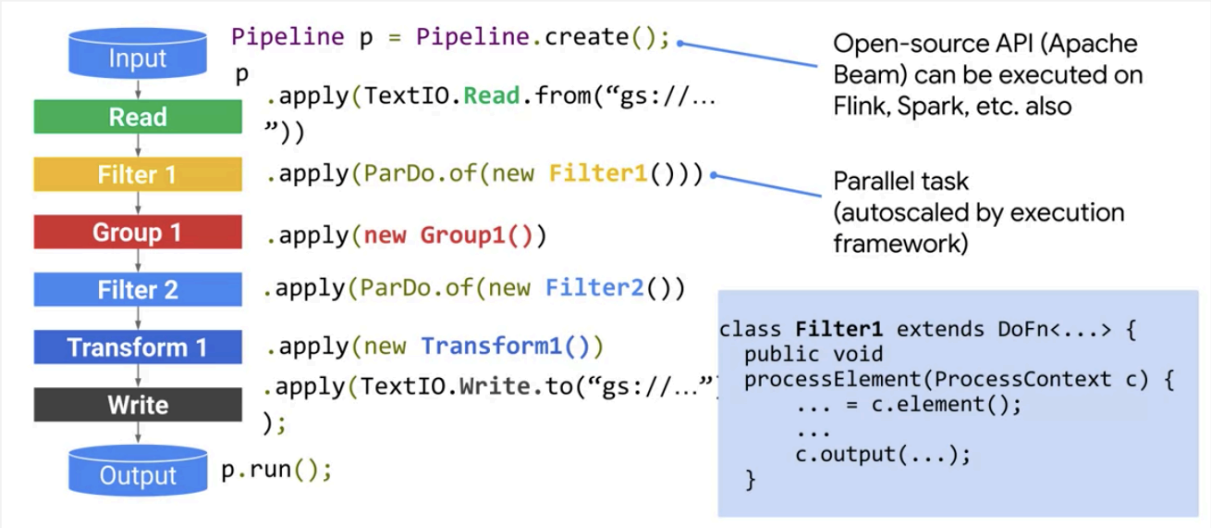

Figure 3. An example of a Pipeline using the Apache Beam SDK

Implementing streaming pipelines on Cloud Dataflow

What is Cloud Dataflow

Could Dataflow is a serverless unified stream and processing of batch data that is fast, and cost-effective?

Automated resource control and rebalancing of complex jobs

Dataflow automates the provision and deployment of computing resources to minimize latency and optimize utilization so that instances don't need to spin up or reserve by hand. Job partitioning is often automated and programmed so that the lagging job is dynamically rebalanced. No need to chase "hotkeys" or preprocess data from your inputs.

Horizontal Autoscaling

Horizontal autoscaling of workers results in better overall price-to-performance for optimal efficiency.

Dynamic resource scheduling for batch processing

Flexible resource scheduling (FlexRS) provides a cheaper price for batch processing to produce more flexibility in task scheduling time, such as overnight work. Such versatile jobs are placed into a queue with a promise that they will be retrieved within a six-hour period for execution.

Workflow with Dataflow

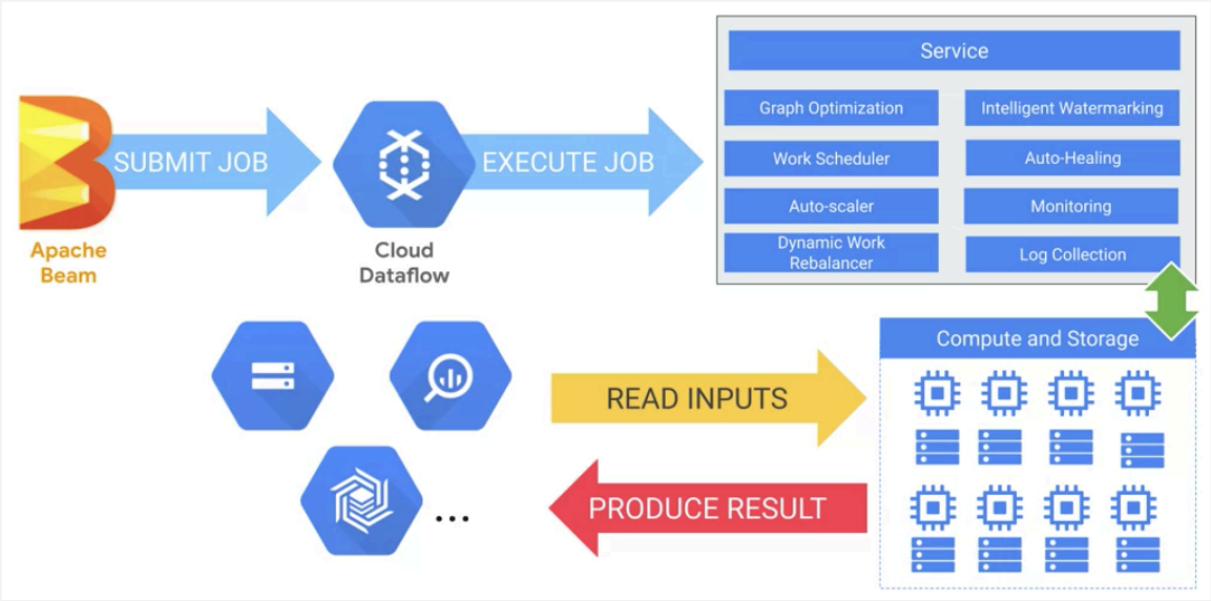

Figure 4. Workflow with Dataflow

The Cloud Dataflow software optimizes the model's execution graph to eliminate inefficiencies. It schedules jobs and the fashion distributed to the workers and scales as required.

In case of failure with those workers, it will auto-heal. This will dynamically re-equilibrate to allow maximum use of the underlying jobs. And to produce a result it must bind to a number of data syncs. BigQuery is only one of the many syncs to which you can upload data, by default, is all the computing and storage resources that Cloud Dataflow manages to suit the streaming data pipeline demand elastically.

Visualizing insights with Data Studio

Data visualization is a valuable method for testing the data before showing the results as well.

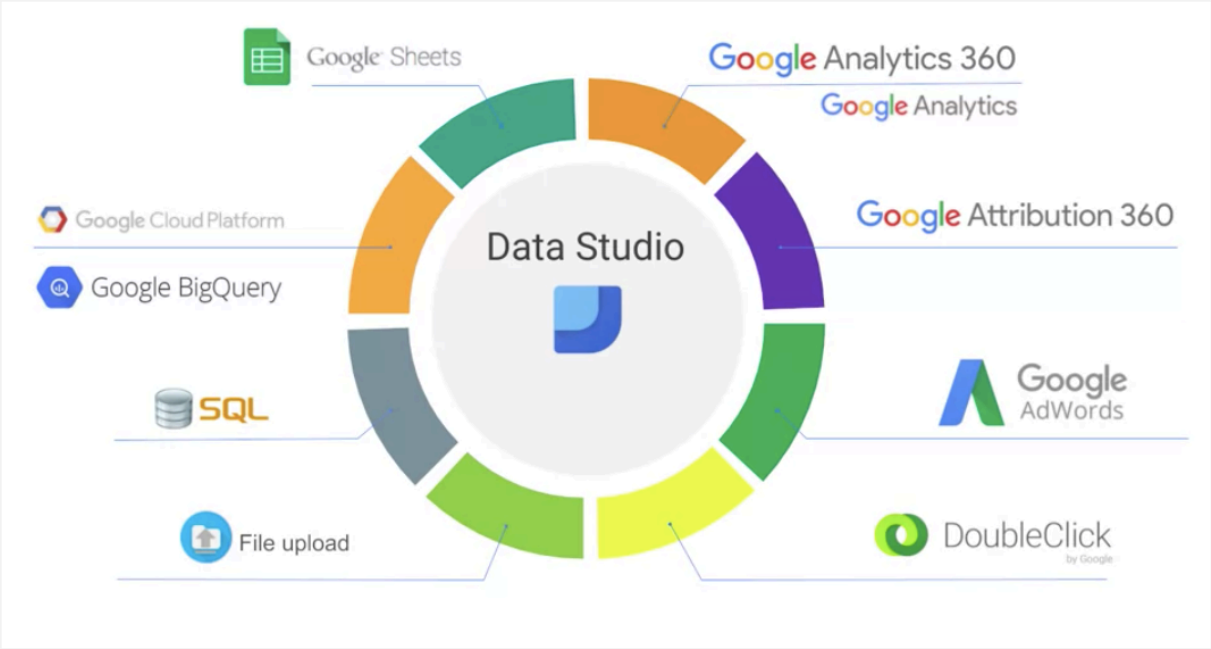

Figure 5. Google Data Studio allows connecting to multiple types of data sources (source: Google Cloud)

References

Apache.org, 2020. *Apache Beam Overview.* [Online]
Available at: https://beam.apache.org/get-started/beam-overview/
[Accessed 2020].

daCosta, F. & Santa, C., 2014. *Rethinking the Internet of Things.* s.l.:s.n.

Google Cloud LLC, 2020. *Google Cloud.* [Online]
Available at: https://cloud.google.com
[Accessed 2020].

Google LLC, 2014. *Better data centers through machine learning.* [Online]
Available at: https://googleblog.blogspot.com/2014/05/better-data-centers-through-machine.html
[Accessed 2020].

Statista, 2020. *Internet of Things (IoT) active device connections installed base worldwide from 2015 to 2025*.* [Online]
Available at: https://www.statista.com/statistics/1101442/iot-number-of-connected-devices-worldwide/
[Accessed 2020].