OCTOBER 5, 2020 / #DATA SCIENCE

# How to Build a Scalable Data Analytics Pipeline

Priyanka Vergadia

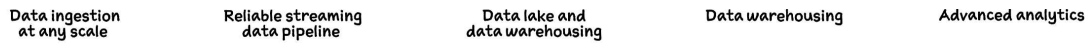
Every application generates data, but what do those data mean? This is a question all data scientists are hired to answer.

There is no doubt that this information is the most precious commodity for a business. But making sense of data, creating insights and turning them into decisions, is even more important.

As the data keep growing in volume, the data analytics pipelines have to be scalable to adapt the rate of change. And for this reason, choosing to set up the pipeline in the cloud makes perfect sense (since the cloud offers on-demand scalability and flexibility).

In this article I will demystify how to build a scalable and adaptable data processing pipeline in Google Cloud. And don't worry – these concepts are applicable in any other cloud or on-premise data pipeline.

## 5 Steps to Create a Data Analytics Pipeline:




5 steps in a data analytics pipeline

- First you ingest the data from the data source
- Then process and enrich the data so your downstream system can utilize them in the format it understands best.
- Then you store the data into a data lake or data warehouse for either long term archival or for reporting and analysis.
- You can then analyze the data by feeding them into analytics tools.
- Apply machine learning for predictions or create reports to share with your teams.

Let's go through each of these steps in more detail.

# How to Capture the data

Depending on where your data is coming from, you can have multiple options to ingest them.

- Use data migration tools to migrate data from on-premises or from one cloud to another. Google Cloud offers a storage transfer service for this purpose.
- To ingest data from your 3rd party saas services, use APIs and send the data to the data warehouse. In Google Cloud BigQuery, the serverless data warehouse provides a data transfer service that allows you to bring in data from saas apps such as YouTube, Google Ads, Amazon S3, Teradata, ResShift and more.

event messages into Pub/Sub from where a subscriber picks up the message and takes appropriate action on it.

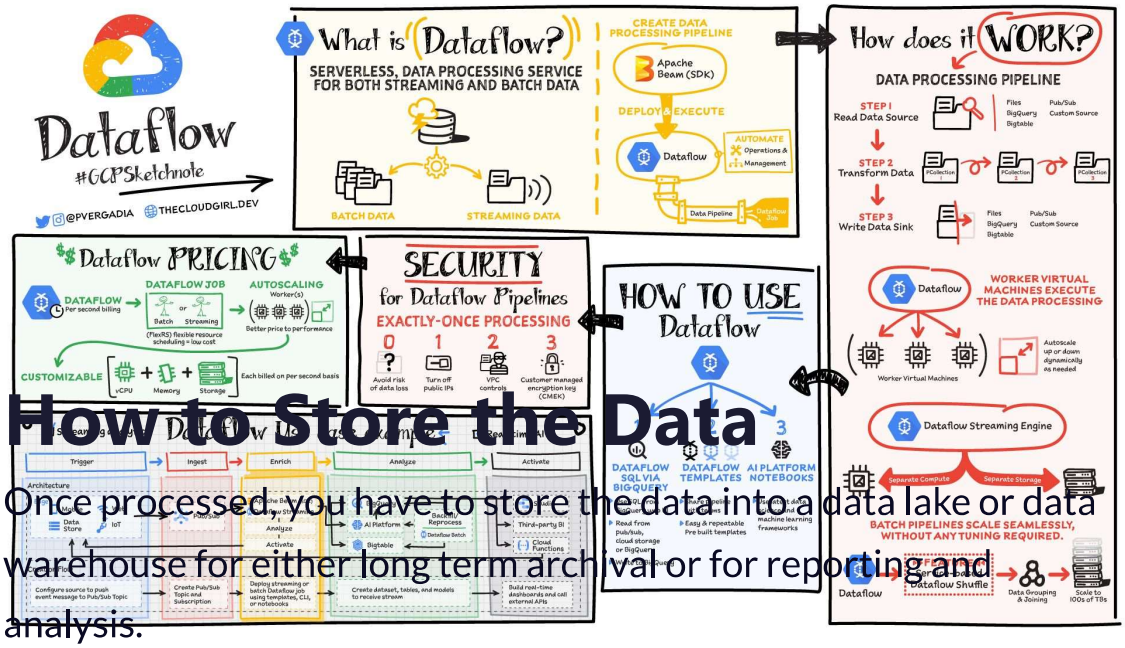
- If you have IoT devices they can stream real-time data using Cloud IoT core which supports MQTT protocol for the IoT devices. You could also send IoT data to Pub/Sub.

# How to Process the Data

Once the data is ingested, they need to be processed or enriched in order to make them useful for the downstream systems.

There are three main tools that help you do that in Google Cloud:

- Dataproc is essentially managed Hadoop. If you use the Hadoop ecosystem then you know that it can be complicated to set it up, involving hours and even days. Dataproc can spin up a cluster in 90 seconds so you can start analyzing the data quickly.
- Dataprep is an intelligent graphical user interface tool that helps data analysts process data quickly without having to write any code.
- Dataflow is serverless data processing service for streaming and batch data. It is based on the Apache Beam open source SDK making your pipelines portable. The service separates storage from computing, which allows it to scale seamlessly. For more details refer to the GCPSketchnote below.




## How to Store the Data

Once processed, you have to store the data into a data lake or data warehouse for either long term archival or for reporting and analysis.

There are two main tools that help you do that in Google Cloud:

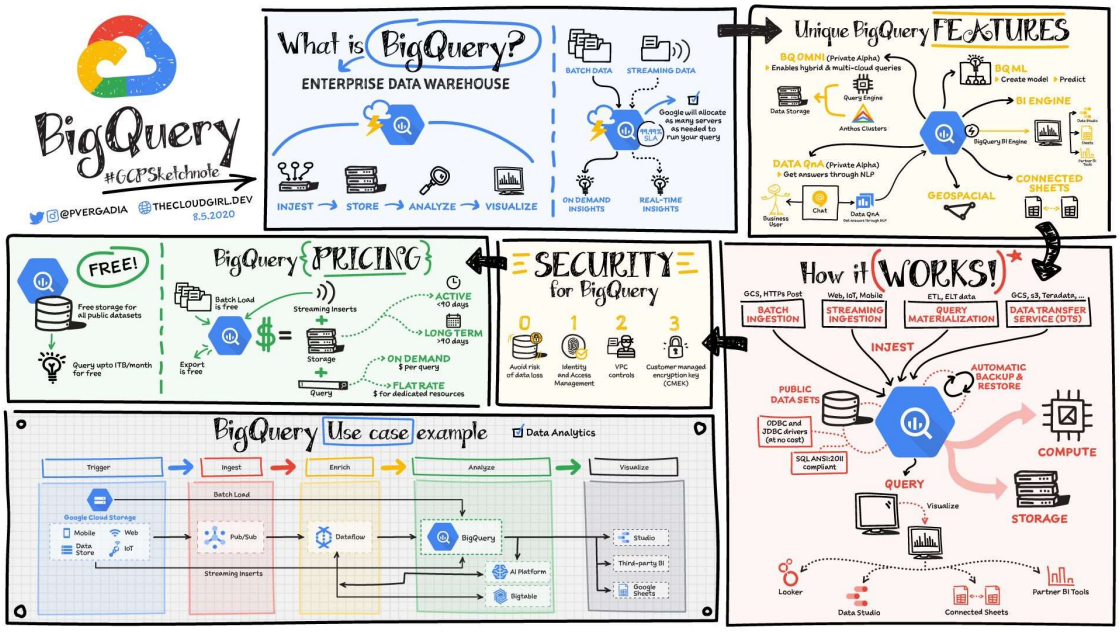
Google Cloud Storage is an object store for images, videos, files and so on which comes in 4 types:

1. Standard Storage: Good for “hot” data that’s accessed frequently, including websites, streaming videos, and mobile apps.
2. Nearline Storage: Low cost. Good for data that can be stored for at least 30 days, including data backup and long-tail multimedia content.
3. Coldline Storage: Very low cost. Good for data that can be stored for at least 90 days, including disaster recovery.
4. Archive Storage: Lowest cost. Good for data that can be stored for at least 365 days, including regulatory archives.

BigQuery is a serverless data warehouse that scales seamlessly to petabytes of data without having to manage or maintain any server.

easily share the data and queries with others on your team.

It also houses 100's of free public datasets that you can use in your analysis. And it provides built-in connectors to other services so data can be easily ingested into it and extracted out of it for visualization or further processing/analysis.




# How to Analyze the Data

Once the data is processed and stored in a data lake or data warehouse, they are ready to be analyzed.

If you are using BigQuery to store the data, then you can directly analyze that data in BigQuery using SQL.

If you are using Google Cloud Storage then you can easily move the data into BigQuery.

BigQuery also offers Machine Learning features with BigQueryML.

# How to Use and Visualize the Data

## Using the data

Once the data are in the data warehouse you can use them to get insights and to make predictions using machine learning.

For further processing and predictions you can use the Tensorflow framework and AI Platform depending on your needs.

Tensorflow is an end-to-end open source machine learning platform with tools, libraries, and community resources.

AI Platform makes it easy for developers, data scientists, and data engineers to streamline their ML workflows. It includes tools for each stage of the ML lifecycle starting from Preparation --> Build --> Validation --> Deployment.

## Visualizing the data

There are lots of different tools for data visualization, and most of them have a connector to BigQuery to easily create charts in the tool of your choice.

Google Cloud provides a few tools that you might find helpful to look at.

- Data Studio is free and connects not just to BigQuery but also to many other services for easy data visualization. If you have used Google Drive, sharing charts and dashboards are exactly like that – extremely easy.

# Conclusion

There is a lot that goes on in a data analytics pipeline. Whichever tools you choose to use, make sure they can scale as your data grow in the future.

For more such content, you can follow me on Twitter, @pvergadia and visit my website, thecloudgirl.dev.

---

**Priyanka Vergadia**

I am currently Developer Advocate at Google where I have created over 300 videos, articles, podcasts, courses and tutorials that have helped developers learn Google Cloud fundamentals.

---

**You can make a tax-deductible donation here.**