Cognizant

Google Cloud Platform

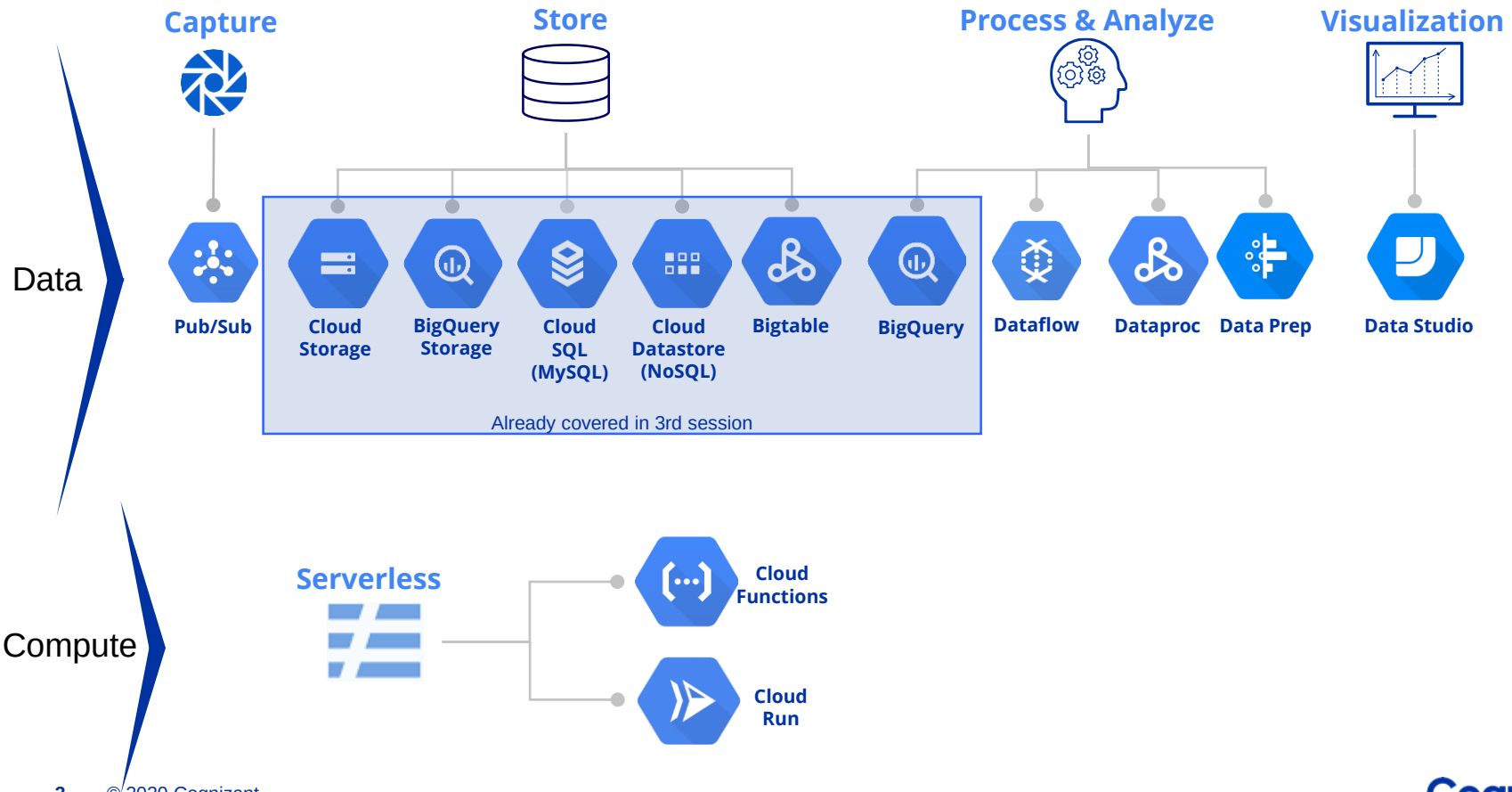
Data and Serverless Services

May 20, 2020

Agenda - Master Class

S.No.	Topic (Master Class)	Date	Day
1	Introduction to Google Cloud Platform	4-May	Mon
2	Introduction to Google Compute Services - GCE | GAE | GKE	6-May	Wed
3	Introduction to Google Storage - GCS | Bigtable | Big Query | Datastore	8-May	Fri
4	Introduction to Google Networking	11-May	Mon
5	Introduction to GCP Monitoring Services	13-May	Wed
6	**DEMO-I (2 Hours)**	15-May	Fri
7	Introduction to GCP Security Services	18-May	Mon
8	***Introduction to Google Data & Serverless Services***	20-May	Wed
9	Introduction to GCP DevOps Services	22-May	Fri
10	Introduction to Google API Services	26-May	Tue
11	Introduction to Google Anthos	27-May	Wed
12	**DEMO-II (2 Hours)**	29-May	Fri

Data and Serverless services



Serverless

Cloud Functions

Cloud Run

Cognizant

Cloud Functions

Overview

- Server-less managed service
- Simple, single-purpose functions attached to events emitted from cloud infrastructure and services

- Triggered when an event being watched is fired
  - HTTP Triggers
  - Cloud Pub/Sub Triggers
  - Cloud Storage Triggers
  - Direct Triggers
  - Cloud Firestore
  - Analytics for Firebase
  - Firebase Realtime Database
  - Firebase Authentication

- Functions can be written in Node.js, Python and Go

Features

- No worry about infrastructure e.g. VM size/count
- Pay-Per-Use
- Deploy functions not apps
- Portable code
- It's event based/oriented
- Scales Independently
- Isolated from one another
- Stateless and Ephemeral

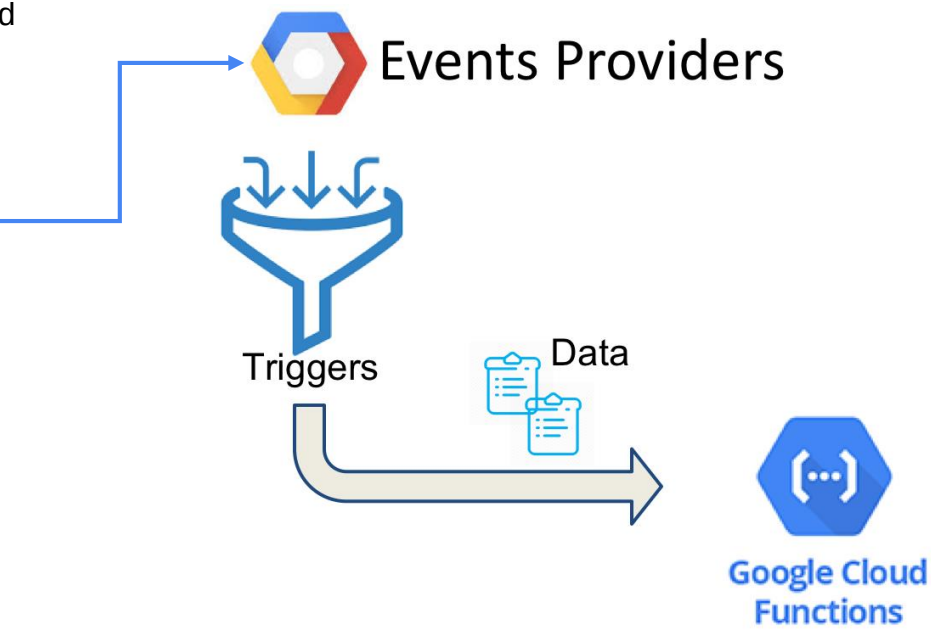


Image Source: Medium.com

Functions - Types

Foreground Functions (HTTP):

- Invoke HTTP functions from standard HTTP requests
- A TLS certificate is automatically provisioned, so all HTTP Functions can be invoked via a secure connection

Background Functions:

- Asynchronously invoked by events
- Handle events from your Cloud infrastructure
  - ✓ Cloud Storage
  - ✓ Cloud Pub/Sub
  - ✓ Firebase

Function - Example (Node.js)

Foreground Functions (HTTP):

```
const escapeHtml = require('escape-html');

/**
 * HTTP Cloud Function.
 *
 * @param {Object} req Cloud Function request context.
 *                     More info: https://expressjs.com/en/api.html#req
 * @param {Object} res Cloud Function response context.
 *                     More info: https://expressjs.com/en/api.html#res
 */
exports.helloHttp = (req, res) => {
  res.send(`Hello ${escapeHtml(req.query.name || req.body.name ||
'World')}!`);
};
```

---

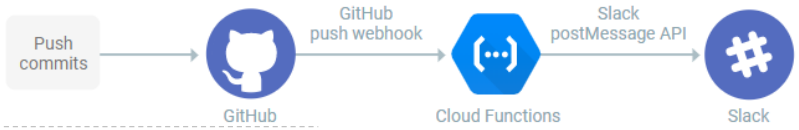
curl -X POST **HTTP_TRIGGER_ENDPOINT** -H "Content-Type:application/json" -d '{"name":"Jane"}'

Background Functions:

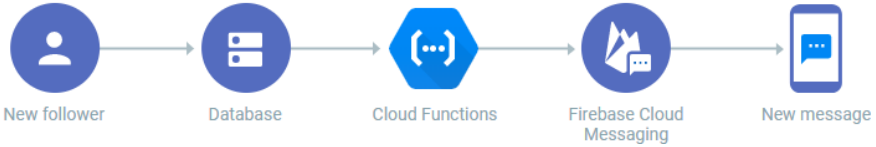
```
/**
 * Background Cloud Function.
 *
 * @param {object} data The event payload.
 * @param {object} context The event metadata.
 */
exports.helloBackground = (data, context) => {
  return `Hello ${data.name || 'World'}!`;
};
```

Use cases

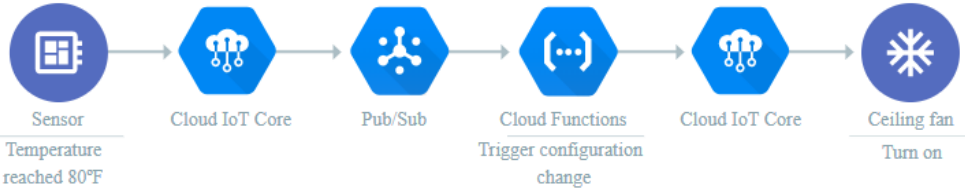
Integration with third-party services and APIs



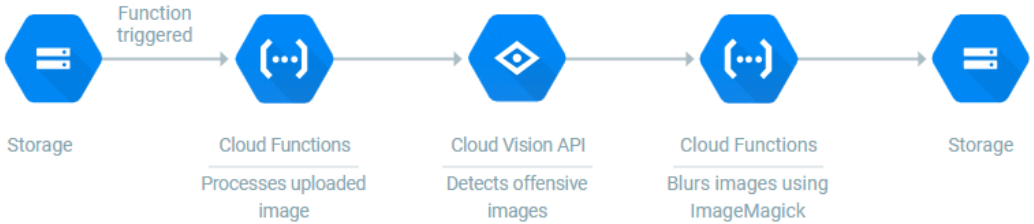
Serverless mobile back ends



Serverless IoT back ends



Real-time stream processing



Functions - Comparison

Feature	AWS Lambda	Google Cloud Functions	Azure Functions
Scalability & availability	Automatic scaling (transparently)	Automatic scaling	Manual or metered scaling (App Service Plan), or sub-second automatic scaling (Consumption Plan)
Max # of functions	Unlimited functions	1,000 functions per project	Unlimited functions
Concurrent executions	600 parallel executions per account, per region (ask to customer service for greater limit)	No limit	No limit
Max execution	300 sec (5 min)	No limit	300 sec (5 min)
Supported languages	JavaScript, Java Python, C#	Only JavaScript	C# and JavaScript (preview of F#, Python, Batch, PHP, PowerShell)
Dependencies	Deployment Packages	npm package.json	Npm, NuGet
Deployments	Only ZIP upload (to Lambda or S3)	ZIP upload, Cloud Storage or Cloud Source Repositories	Visual Studio Team Services, OneDrive, Local Git repository, GitHub, Bitbucket, Dropbox, External repository
Environment variables	Yes	Not yet	App Settings and ConnectionStrings from App Services
Versioning	Versions and aliases	Cloud Source branch/tag	Cloud Source branch/tag
Event-driven	S3, SNS, SES, DynamoDB, Kinesis, CloudWatch, Cognito, API Gateway, CodeCommit, etc.	Cloud Pub/Sub or Cloud Storage Object Change Notifications	Blob, EventHub, Generic WebHook, GitHub WebHook, Queue, Http, ServiceBus Queue, Service Bus Topic, Timer triggers
HTTP(S) invocation	API Gateway	HTTP trigger	HTTP trigger
Logs management	CloudWatch	Cloud Logging	App Services monitoring
In-browser code editor	Yes	Only with Cloud Source Repositories	Functions environment, AppServices editor
Granular IAM	IAM roles	IAM Roles	IAM roles
Pricing	1M requests for free (Free Tier), then $0.20/1M requests	2M requests for free (Free Tier), then $0.40(charged at a per-unit rate of $0.0000004 per invocation)	1 million requests and 400,000 GB-s (Free Grant), then $0.000016/GB-s per execution time and $0.20/1M executions

Cloud Run

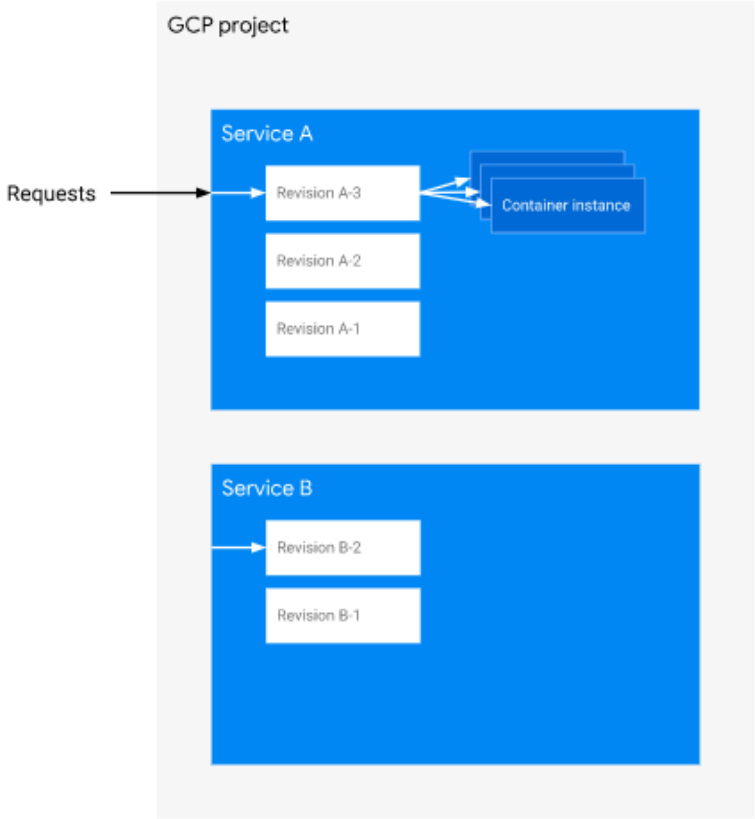
Overview

- Fully managed compute platform for deploying and scaling **stateless** containerized applications quickly and securely
- Cloud Run is built on the Knative open source project, enabling portability of your workloads across platforms

- The **service** is the main resource of Cloud Run
- Each deployment to a service creates a **revision**
- Each revision receiving requests is automatically scaled to the number of container instances needed to handle all these requests
- Supports version based traffic routing. Useful for Canary deployments and Rollbacks
- Services can be deployed to fully managed platform or **Anthos** GKE cluster

Competitive Products

- AWS Fargate on EKS
- Azure Container Instances (ACI)



Competition

Capability / Feature	AWS Fargate on EKS	Google Cloud Run	Azure Container Instances
Deployment Unit	Kubernetes Pod	Container Image	Container Image
Isolation Level	Dedicated VM	gVisor	Dedicated VM
Windows Containers	No	No	Yes
Deployment Spec	Kubernetes Pod	Knative Service	ACI Native Spec
Multiple Containers	Yes	No	Yes
Native Persistence	None	None	Yes (Azure File Share)
Orchestration Support	EKS / ECS	Anthos	AKS (Virtual Kubelet)
Cluster Pre-provisioning	Required (EKS)	Optional (Anthos)	Not Required
Terraform Support	Yes	Yes	Yes
GPU Support	No	Yes (Anthos)	Yes (Preview)
Public IP / CNAME	No	Yes	Yes
In-Built Auto Scaling	Yes (EKS + HPA)	Yes	Yes (AKS + HPA)
Scale-to-Zero	Yes	Yes	Yes
Virtual Network Access	Yes	Yes	Yes
Logging	CloudWatch	StackDriver	Azure Monitor Logs
Revisions / Versioning	No	Yes	No

Source: [The New Stack]

Console

Google Cloud Platform | Cloud academy-1000184377

Cloud Run | Create service

1 Service settings

Deployment platform and service name are the identifier of a service; they can't be changed once deployed.

Deployment platform

- (●) Cloud Run (fully managed)

  Region *
  us-central1 (Iowa)
  How to pick a region?

- ( ) Cloud Run for Anthos

Service name *

Service name

Authentication *

- ( ) Allow unauthenticated invocations
  Check this if you are creating a public API or website.
- ( ) Require authentication
  Manage authorized users with Cloud IAM.

NEXT

2 Configure the service's first revision

CANCEL

Google Cloud Platform | Cloud academy-1000184377

Cloud Run | Create service

✓ Service settings

2 Configure the service's first revision

A service can have multiple revisions. The configurations of each revision are immutable.

Container image URL *
gcr.io/cloudrun/hello SELECT

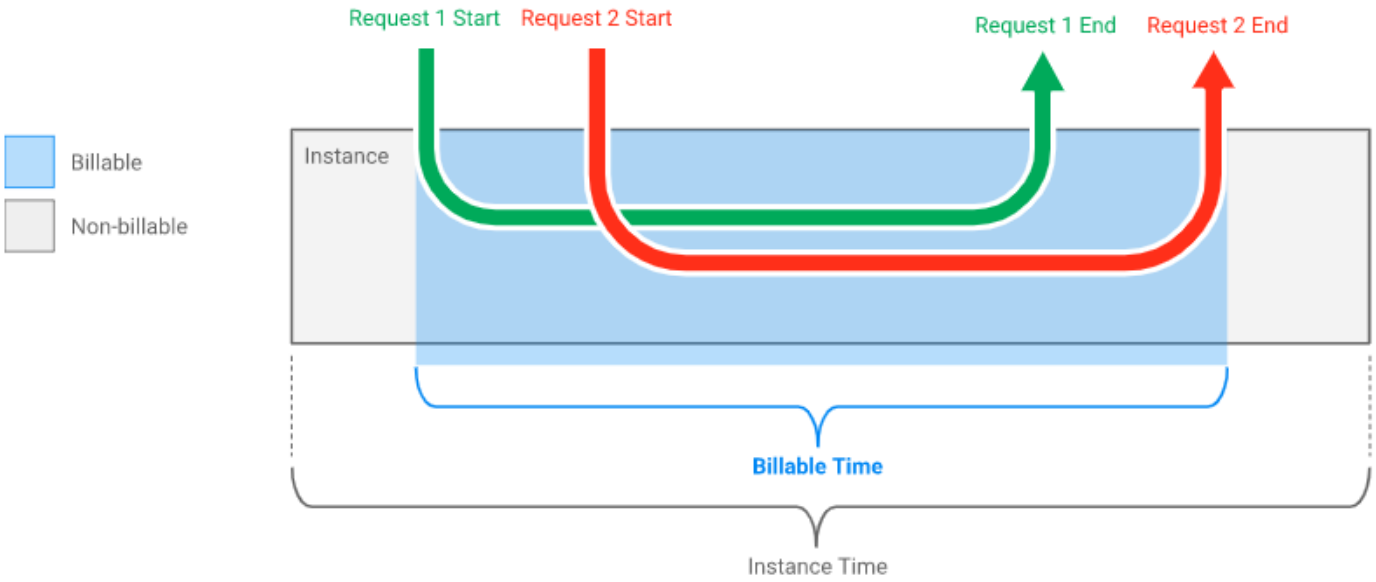
E.g. gcr.io/cloudrun/hello
Should listen for HTTP requests on $PORT and not rely on local state. How to build a container?

SHOW ADVANCED SETTINGS

CREATE CANCEL

Cognizant

Pricing



CPU	Memory	Requests	Networking
First 180,000 vCPU-seconds free	First 360,000 GiB-seconds free	2 million requests free	1 GiB free egress within North America
$0.00002400 / vCPU-seconds beyond free quota	$0.00000250 / GiB-second beyond free quota	$0.40 / million requests beyond free quota	Google Cloud Network Premium tier pricing beyond free quota.

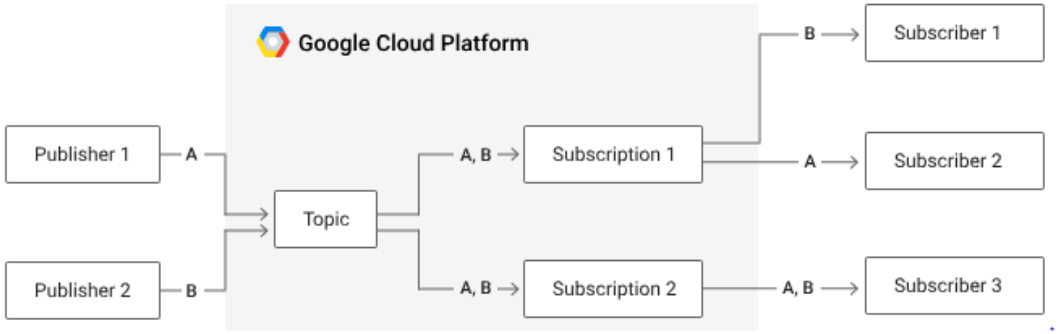
Data Capture

Cloud Pub/Sub

Pub/Sub - Basics

There are several key concepts in a Pub/Sub service:

- **Message:** the data that moves through the service.
- **Topic:** a named entity that represents a feed of messages
- **Subscription**: a named entity that represents an interest in receiving messages on a particular topic.
- **Publisher (also called a producer):** creates messages and sends (publishes) them to the messaging service on a specified topic
- **Subscriber (also called a consumer)**: receives messages on a specified subscription



Competitive Products: AWS SNS, Azure Service Bus

Cloud Pub/Sub

Asynchronous many-to-many messaging service from Google

Features

Fully managed

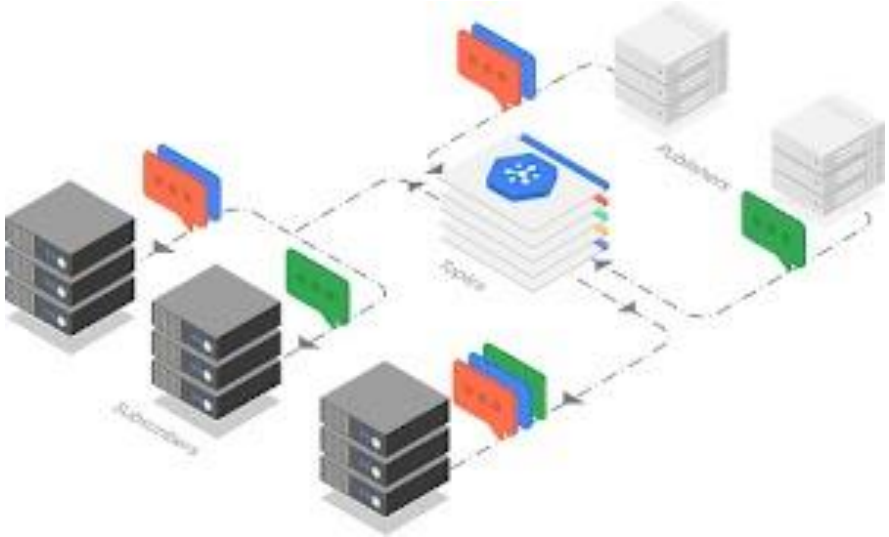
At-least-once delivery

Exactly-once processing

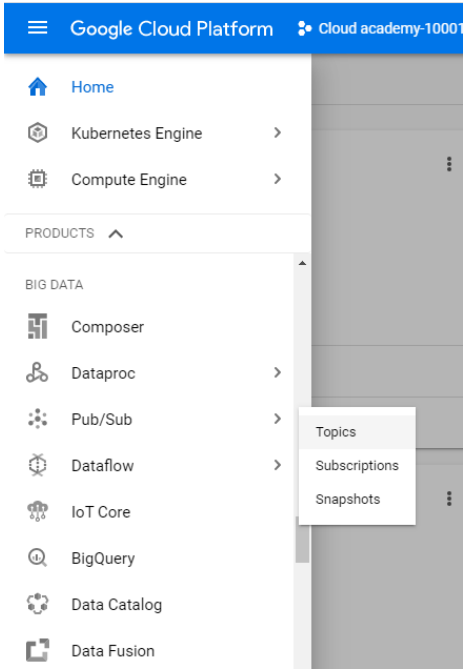
Global by default

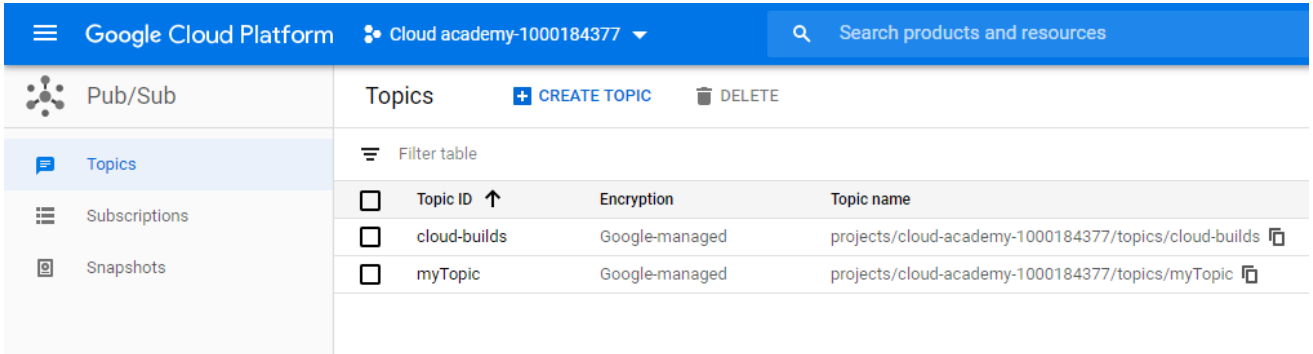
Seek and replay

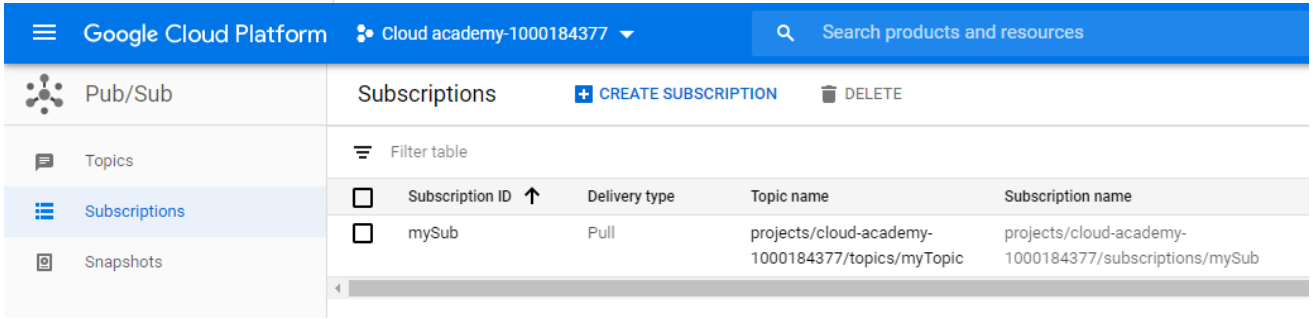
Open APIs and client libraries



Pub/Sub on Console







Pricing

Monthly data volume[1]	Price Per TB[2]
First 10 GB	$0.00
Beyond 10 GB	$40

[1] For detailed pricing information, please consult the pricing guide.
[2] TB refers to a **tebibyte**, or 2^{40} bytes. 1 Tebibyte = 1.1 Terabyte
If you pay in a currency other than USD, the prices listed in your currency on Cloud Platform SKUs apply.

Sample – Publisher and Subscriber (Python)

```python
from google.cloud import pubsub_v1


publisher = pubsub_v1.PublisherClient()

topic_path = publisher.topic_path(project_id, topic_name)

for n in range(1, 10):
    data = u"Message number {}".format(n)

    data = data.encode("utf-8")

    future = publisher.publish(topic_path, data=data)
    print(future.result())

print("Published messages.")
```

1

2

3

```python
from google.cloud import pubsub_v1


subscriber = pubsub_v1.SubscriberClient()
subscription_path = subscriber.subscription_path(
    project_id, subscription_name
)

def callback(message):
    print("Received message: {}".format(message))
    message.ack()

streaming_pull_future = subscriber.subscribe(
    subscription_path, callback=callback
)
print("Listening for messages on {}..\n".format(subscription_path))

with subscriber:
    try:
        streaming_pull_future.result(timeout=timeout)
    except:  # noqa
        streaming_pull_future.cancel()
```

4

5

6

Data Processing

Cloud Dataproc

Cloud Dataflow

Cloud Dataprep

Cognizant

Data Transformation and Processing

Data from source systems is cleansed, normalized, and processed across multiple machines and stored in analytical systems.

Cloud Dataproc

- Existing Hadoop/Spark Applications
- Machine Learning / Data Science Ecosystem
- Tunable Cluster Parameters

Cloud Dataflow

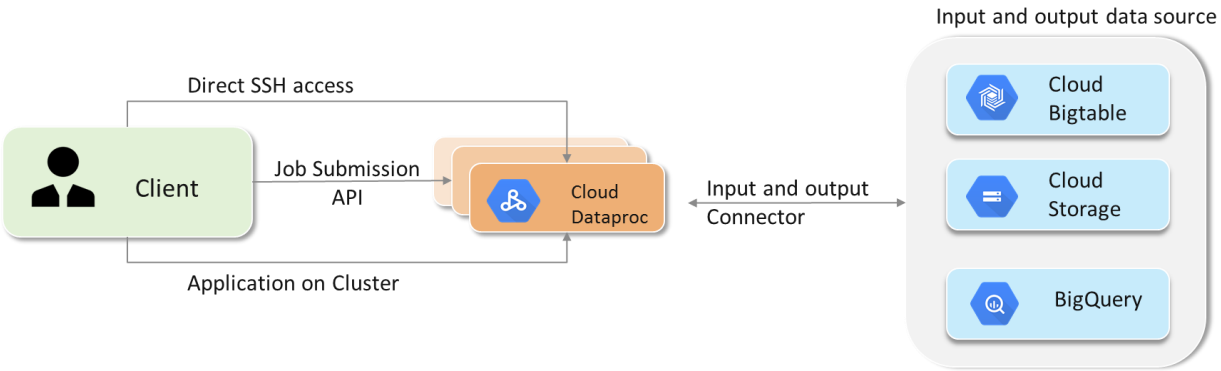
- New Data Processing Pipelines
- Unified Streaming & Batch
- Fully-Managed, No-Ops

Cloud Dataprep

- UI-Driven Data Preparation
- Scales On-Demand
- Fully-Managed, No-Ops

Cloud Dataproc

It's a managed Apache Spark and Apache Hadoop service that lets you take advantage of open source data tools for batch processing, querying, streaming and machine learning



Cloud Dataproc can integrate seamlessly with following GCP services: GCE, GCS, Google BigQuery & Cloud Bigtable thereby providing a powerful and complete data processing platform

Competitive Products: AWS EMR, Azure HDInsight

Dataproc - Connectors

We have different connectors for Dataproc for integration with Google services:

BigQuery Connector: enable programmatic read/write access to Google BigQuery. This is an ideal way to process data that is stored in BigQuery

Cloud Storage Connector: The GCS connector helps running Apache Hadoop or Apache Spark jobs directly on data in Cloud Storage and better than choosing the Hadoop Distributed File System (HDFS)

Bigtable Connector: The GCS connector helps running Apache Hadoop or Apache Spark jobs directly on data in Bigtable which is capable of handling massive streaming data.

Dataproc - Competitive Advantages

Product characteristics	Cloud Dataproc	Competitors	Customer benefit
Cluster start time Elapsed time from cluster creation until it's ready	< 90 seconds	5-30 minutes*	Faster data processing workflows, because less time is spent waiting for clusters to provision and start executing applications.
Billing unit of measure Increment used for billing service when active	Minute	Hourly	Reduced costs for running Spark and Hadoop; you pay for what you actually use, not a cost that's been rounded up.
Preemptible VMs Clusters can utilize preemptible VMs	Yes	No	Lower total operating costs for Spark and Hadoop processing by leveraging the cost benefits of preemptible VMs.
Job output and cancellation Jobs are cancelable and output is easy to find	Yes	No	Higher productivity; job output doesn't necessitate reviewing log files and canceling jobs. Doesn't require SSH.
Custom machine types Size the machine to the job.	Yes	No	Get the exact amount of processing power and memory you need. Don't buy excess.

Cloud Dataflow

Tool for developing & executing data processing patterns e.g. Extract, Transform and Load (ETL), on very large data sets

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

Any kind of data processing task, encompassing both batch and streaming data processing

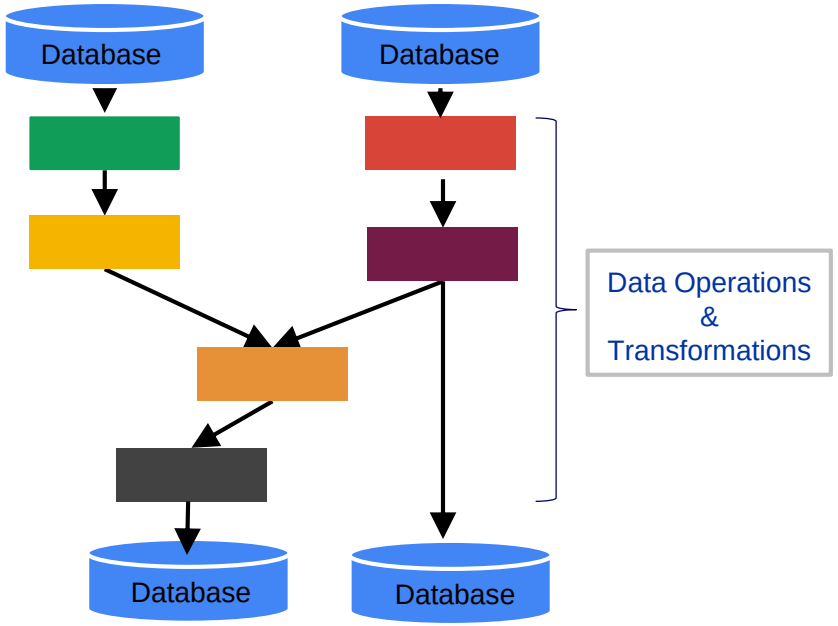
Dataflow can process practically any amount of data arriving at regular/irregular intervals

Dataflow is particularly useful for embarrassingly parallel data processing tasks, in which the problem can be decomposed into many smaller bundles of data that can be processed independently, making it very fast

Competitive Products: AWS Kinesis, Azure Stream Analytics

What are data pipelines?

- A data processing pipeline is a program that is used for completing a data processing job
- Defined as a set of data processing transformations
- Optimized and executed as a unit
- Can include multiple inputs and multiple outputs
- Can perform many mathematical, logical, or transformation operations and might include filtering, grouping, comparing, or joining data
- PCollections (a distributed data set) conceptually flows through the pipeline



Google Dataflow - Features

Pipeline first, runtime second - With the Dataflow model and SDKs, first focus is on defining data pipelines, not how they'll run or the characteristics of the particular runner executing them.

Portability - Data pipelines are portable across a number of runtime engines. You can choose a runtime based on any number of considerations, such as performance, cost or scalability.

Development tooling - The Dataflow SDK contains the tools to create portable data pipelines using open-source languages, libraries and tools

Dynamic Work Rebalancing: Automated and optimized work partitioning dynamically rebalances lagging work

Unified model - Batch and streaming are integrated into a unified model with windowing, ordering and triggering.

Automated Resource Management: Automates provisioning & management of processing resources to minimize latency and maximize utilization

Horizontal Auto-Scaling: Auto-scaling of worker resources for optimum throughput results in better overall price-to-performance

Monitoring: Stackdriver unified logging and monitoring solution, lets you monitor and troubleshoot your pipelines as they are running

Reliable and Consistent Processing: Provides built-in support for fault-tolerant execution that is consistent regardless of data size, cluster size, processing pattern or pipeline complexity

Unified Programming model: Apache Beam SDK offers equally rich MapReduce-like operations, powerful data windowing, and fine-grained correctness control for streaming and batch data alike.

Cloud Dataflow - Use Cases

Batch Data Movement

Moving at rest data from one system to another, such as from Google Cloud Storage to BigQuery

Data reduction & Enrichment

Reduce, compress, re-shape existing data into smaller, computed values, such as log files and geo tags

Continuous Computation

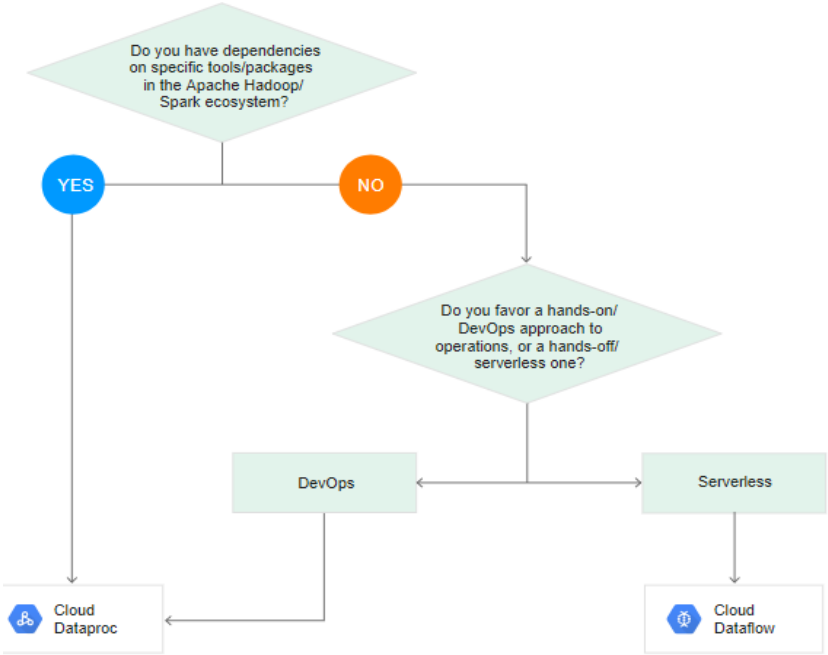
Analyze real-time streaming inputs, such as Click streams

Continuous Data Movement

Real-time ETL over streaming inputs

Cloud Dataproc vs Cloud Dataflow

WORKLOAD	CLOUD DATAPROC	CLOUD DATAFLOW
Stream processing (ETL)		Yes
Batch processing (ETL)	Yes	Yes
Iterative processing and notebooks	Yes	
Machine Learning with Spark ML	Yes	
Pre-processing for machine learning		Yes (with Cloud ML Engine)



What is Dataprep?

Cloud Dataprep by Trifacta™ is an intelligent data service for visually exploring, cleaning and preparing structured and unstructured data for analysis, reporting and machine learning.

Because Cloud Dataprep is serverless and works at any scale, there is no infrastructure to deploy or manage.

The next ideal data transformation is suggested and predicted with each UI input, so you don't have to write code.

Competitive Products: Stitch – A Talend Company

Dataprep - Features

Predictive transformations	Rich transformations	Parameterization	Collaboration
Pattern matching	Standardization	Active profiling	Sampling
Scheduling	Rapid targeting	Common data types	Integrated with Google Cloud Platform

Data Visualization

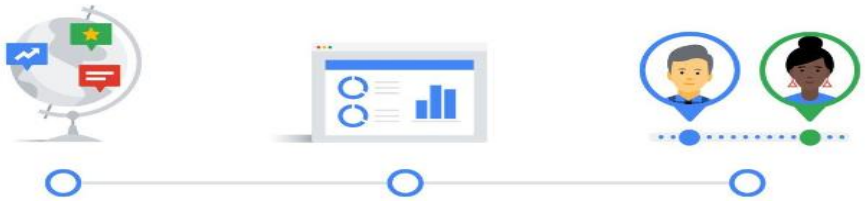
Data Studio

Cognizant

Google Data Studio

What it is?

A **Visualization** tool from Google, available free of cost

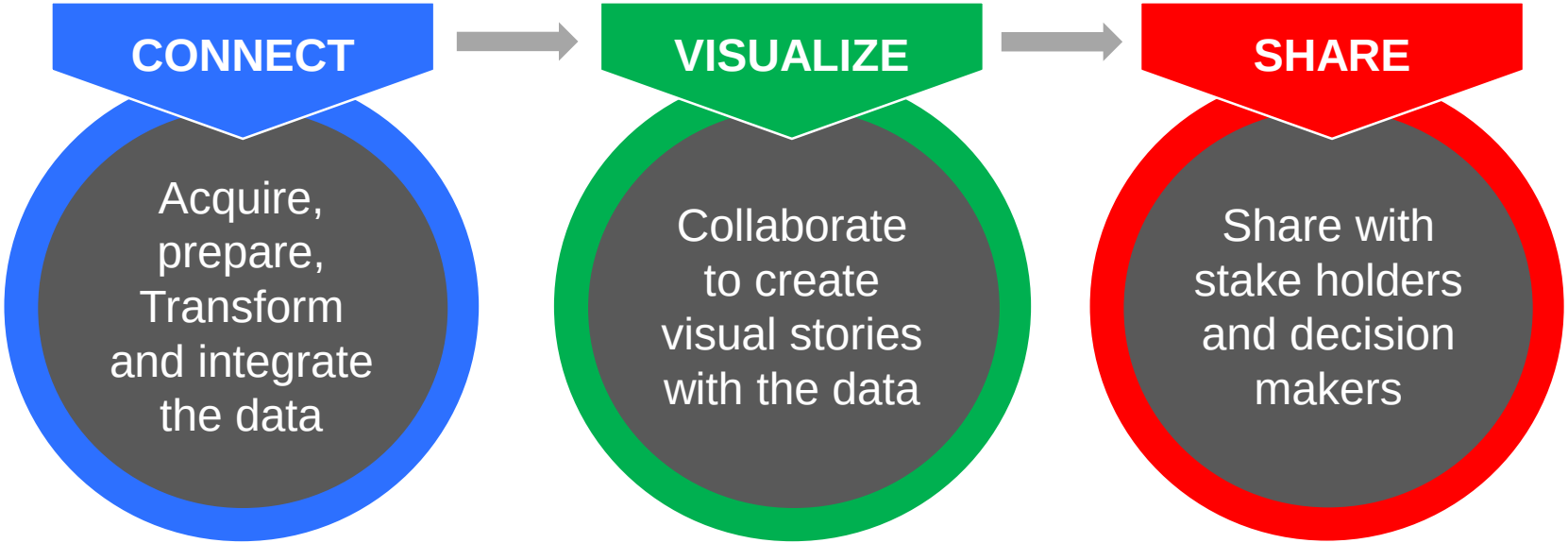


What it does?

It turns your data into informative **Dashboards** and **interactive reports** that are easy to **read**, easy to **share** and fully **customizable**

Data Studio

How it Does?



Data Studio - Features

Unite your data in one place.
Easily connect your data from spreadsheets, Analytics, Google Ads, Google Big Query and more.

Explore the data.
Transform your raw data into the metrics and dimensions needed to create easy-to-understand reports and dashboards — no code or queries required.

Tell impactful stories.
Create and share engaging reports and data visualizations that tell the story for you.

Data Studio Connectors

Google Connectors (17)

Connectors built and supported by Data Studio. Learn more

Google Analytics
By Google
Connect to Google Analytics reporting views. Learn more

Google Ads
By Google
Connect to Google Ads performance report data. Learn more

Google Sheets
By Google
Connect to Google Sheets. Learn more

BigQuery
By Google
Connect to BigQuery tables and custom queries. Learn more

File Upload
By Google
Connect to CSV (comma-separated values) files. Learn more

Campaign Manager
By Google
Connect to Campaign Manager data. Learn more

Cloud Spanner
By Google
Connect to Google Cloud Spanner databases.

Cloud SQL for MySQL
By Google
Connect to Google Cloud SQL for MySQL databases. Learn more

Display & Video 360
By Google
Connect to Display & Video 360 report data.

Extract Data
By Google
Connect to Extract Data Learn more

Google Ad Manager 360
By Google
Connect to Google Ad Manager data. Learn more

Google Cloud Storage
By Google
See your files in Google Cloud Storage. Learn more

MySQL
By Google
Connect to MySQL databases. Learn more

PostgreSQL
By Google
Connect to PostgreSQL databases. Learn more

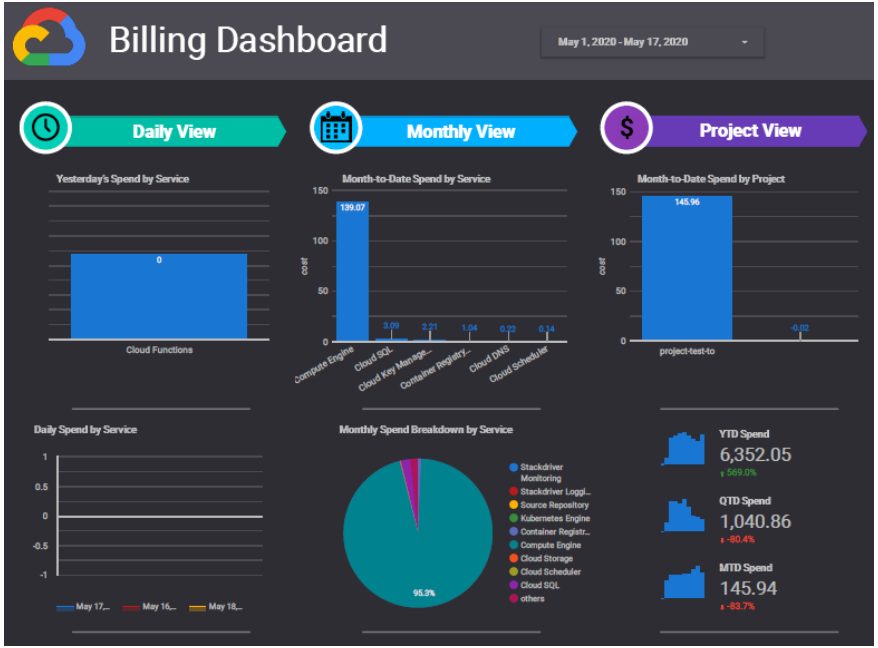
Search Ads 360
By Google
Connect to Search Ads 360 performance reports.

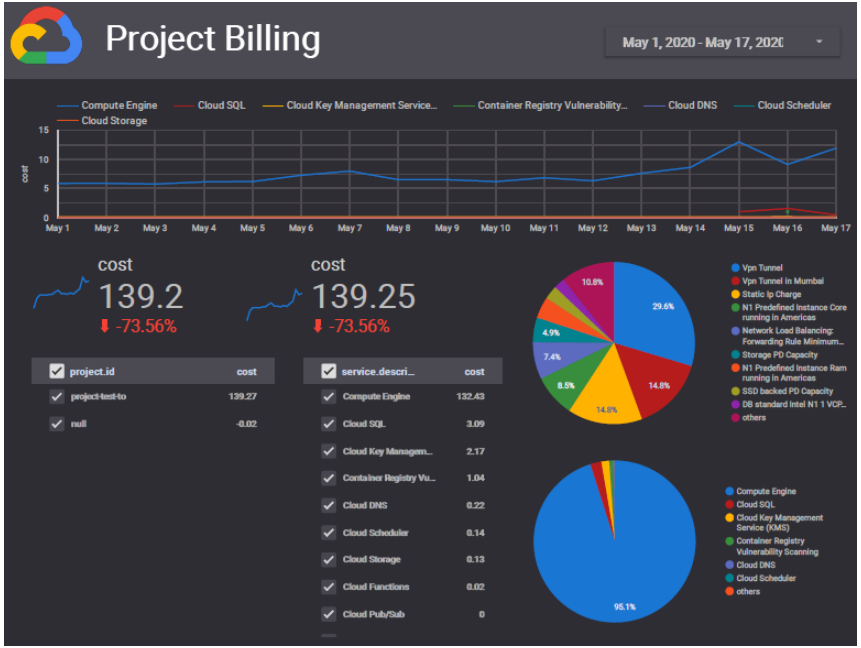
Search Console
By Google
Connect to Search Console data. Learn more

YouTube Analytics
By Google
Connect to YouTube Analytics data. Learn more

https://datastudio.google.com/u/0/datasources/create

Data Studio - Sample Report





References

- https://cloud.google.com/functions/docs/
- https://cloudplatform.googleblog.com/2017/03/Google-Cloud-Functions-a-serverless-environment-to-build-_and-connect-cloud-services_13.html
- https://cloud.google.com/dataproc/
- https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions
- https://cloud.google.com/functions/docs/
- https://cloudplatform.googleblog.com/2017/03/Google-Cloud-Functions-a-serverless-environment-to-build-_and-connect-cloud-services_13.html
- https://www.youtube.com/watch?v=kXk78ihBpiQ
- https://cloud.google.com/pubsub/docs/
- https://cloud.google.com/run/pricing
- https://cloud.google.com/run
- https://thenewstack.io/comparison-aws-fargate-vs-google-cloud-run-vs-azure-container-instances/
- https://www.stitchdata.com/vs/google-cloud-dataprep/

Cognizant

Thank You

Arun Kumar
475634
Arun.kumar311b80@cognizant.com

Cognizant