Data stack for startup

# Introduction

Building a data driven environment is important for modern businesses. As the SaaS and cloud compute getting popular, there are more and more options to build data ecosystems. Among the various options, implementing a cost-efficient solution could be more beneficial than a powerful solution to small businesses or startup. Hence, this article is going to propose a simple data ecosystem using Google Cloud Platform (GCP) and open-source tools.

# Architecture

There are three key components in the proposed ecosystem, which are storage, analytic, and report. To connect each feature, cloud run and other GCP serverless products will be used. Since Google provides discounts to many of their products, it is very easy to control and minimize the monthly costs.

## Storage

Storage is the foundation of the entire ecosystem. It collects raw data and transforms into severable information supporting analytics or reports. To develop the storage feature, we must (1) build data in-flow pipelines, (2) design data schema, and (3) manage and monitor the data quality.

To build the data pipeline, it is necessary to understand the data volume and immediacy. The general guidelines are that the small data volume can implement cloud run or cloud function, while the big volume can use data proc or spark. The low immediacy data can also use cloud run, but high immediacy data might have to use pub/sub or another message queue.

After loading data into the storage, the raw data must be transformed into valuable information. The most important concept in this stage is to level the data. For example, we can set three levels of data: raw, research, and serving. Raw level represents the original data. This type of data usually needs feature engineering to obtain insights. Research level is the experimental features. In many cases, analysts test different hypothesis and make up various metrics; however, the definitions or hypotheses could be changed. As a result, this level serves informative data but still under experiment. The serving data is the finalized version, and it should be the most stable and informative data.

Continuously improving the data quality and monitoring the data pipeline are critical because low quality of data pipeline is unreliable. To managing the data pipeline, we suggest that the developer should explicitly designed tests to ensure data pipeline and data quality. Each test can be host on cloud run invoke before any data in-flow or transformation.

## Analytics

Analytics is a process to generate insights, informative metrics, or algorithms. To make the analytics effective, the ecosystem has to concentrate on reproducibility, deployment, and computing power efficiency during the design stage.

Reproducibility