Name: Subrat Pradhan

E-Mail: [Subrat2804@gmail.com](mailto:Subrat2804@gmail.com)

Data Engineer Assignment

**Question 1:**

**Goggle Pub/Sub** is designed to provide reliable, one-to-many, many-to-one, many-to-many, asynchronous messaging between applications. Where publisher application can send message to a topic and other applications can subscribe that topic to receive messages.

As per your described scenario I will utilize Google’s Pub/Sub to design a data pipeline for process and analysing the real time user activity data from different social platform in following steps:

**Create a Topic:** At first, I have to create a topic which will help to collect user activity data by receiving all messages related to user activity.

**Publishing Messages:** Publish useractivity data to created topic using Pub/Sub client library.

**Process Data:** The next step is to clean user activity data using any tool or Goggle Cloud Platform (GCP) for dataflow pipeline.

**Transform Data:** Within the data flow pipeline, I will apply various on the cleaned data to extract more meaningful information.

**Data Analysis:** After data transformation I will analyse the data according to extract insights as per requirements.

**Store Data:**  Store the processed data for feature use in a database.

**Subscribers:** Send message to subscriber using Pub/Sub which allows to have one or more subscriber to consume user activity data from topic.

**Question 2:**

Goggle’s Data Fusion is a powerful tool that can greatly simplify building and management of data pipelines in data engineering workflow. It helps in transformation and data processing.

1. Data Fusion provides a drag and drop interface which allows to visually design and configure data pipelines. By using this tool there is no need to write extra complex code for visualisation.
2. It also provides a rich set of built-in transformations that can be applied to data in dataflow pipeline. It helps in data cleaning and data normalization operation.
3. By using Data Fusion, it is easy to handle a large amount of dataset and automatically scales resources according to demand.
4. It helps in monitoring the dataflow pipeline work progress. In case any failures or errors, Data Fusion provides detailed error log and allows for easy debugging and resolving issues.
5. It helps to store processed data in Goggle cloud and Big Query for data analysis and feature use. It optimizes the dataflow pipeline process time.

**Question 3:**

Log Processing plays a critical role in data engineering workflow. It offers a central and flexible way of extracting value from those raw log lines. For example, you can extract numerical values from log line with Log Processing, turn these into metrics on Dynatrace Platform, and include them in dashboards and problem detection.

1. Log processing offers real-time monitoring of data pipelines that handle vehicle log data. The company can monitor health performance of data pipelines by analysing log entries.
2. Log processing helps in debugging data pipelines by providing some valuable information. If any error occurs log shows troubleshooting. After that data engineers can identify the root cause and solve that problem.
3. It optimizes data pipelines by identifying bottlenecks and suboptimal performance. The data engineer can identify areas of optimization by optimizing data transformations and turning hardware specification.