Day 11

Topics

1. Security - KMS
2. ETL
   a. Cloud Pub Sub
   b. Cloud Dataproc
   c. Cloud Dataflow
3. Vertex AI
   a. Agent builder

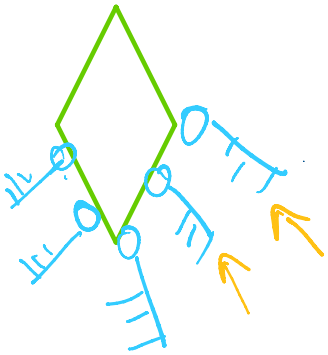
GCP KMS

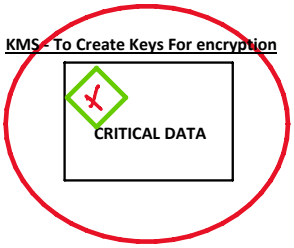
- Key Management Service
- Create, manage and use encryption keys to protect your data.
- GCS, Compute, BQ and Cloud SQL

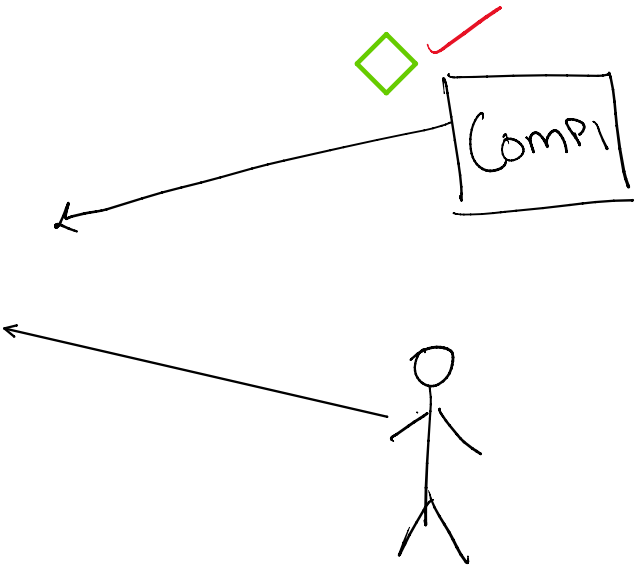
Demo :

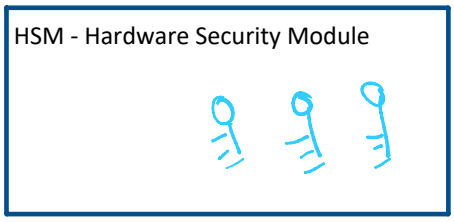
1. Create a KMS key.
2. Use the key to encrypt a GCS.
3. Upload and verify an encrypted file.
4. Simulate a key disablement to test the encryption.

Key Ring
Collection of encryption keys









ETL - Extract, Transform and Load

	Meaning	Example
E - Extract	Collect data from various sources - DB, APIs, Files or any cloud services.	Pull sales data from CSQL. Pull customer data from salesforce.
T - Transform	Clean, filter, aggregate and convert data into a usable format.	Combine sales and customer data, remove duplicates and calculate revenue.
L -Load	Store the transformed data into a destination. - BQ	Save the sales data to BQ or Snowflake [reporting.] Looker - visualization.

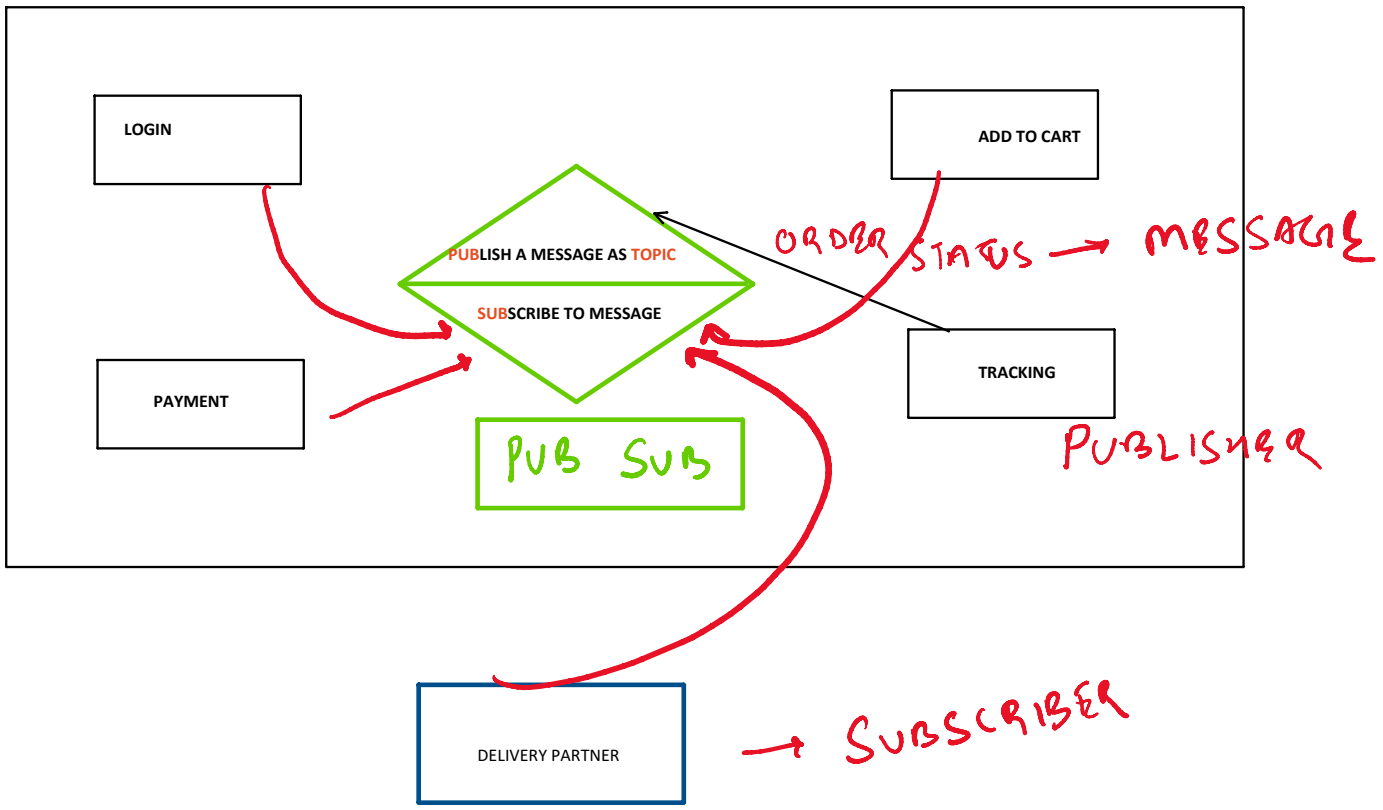
	- Looker - Third party data viewer.	

ETL on Google Cloud

Extract - Pull data using Cloud Pub Sub or Storage
Transform - Process data with Dataflow [Apache Beam] and Dataproc [Hadoop Spark]
Load - Load the data in Big Query.

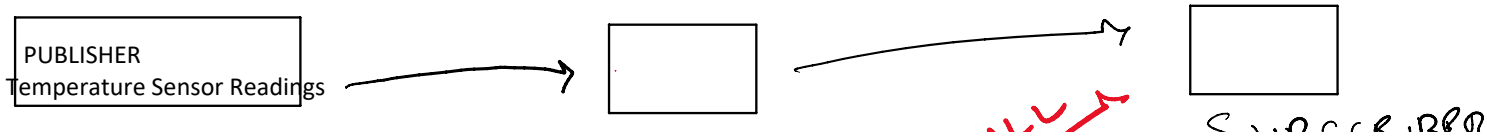
Cloud PUB SUB

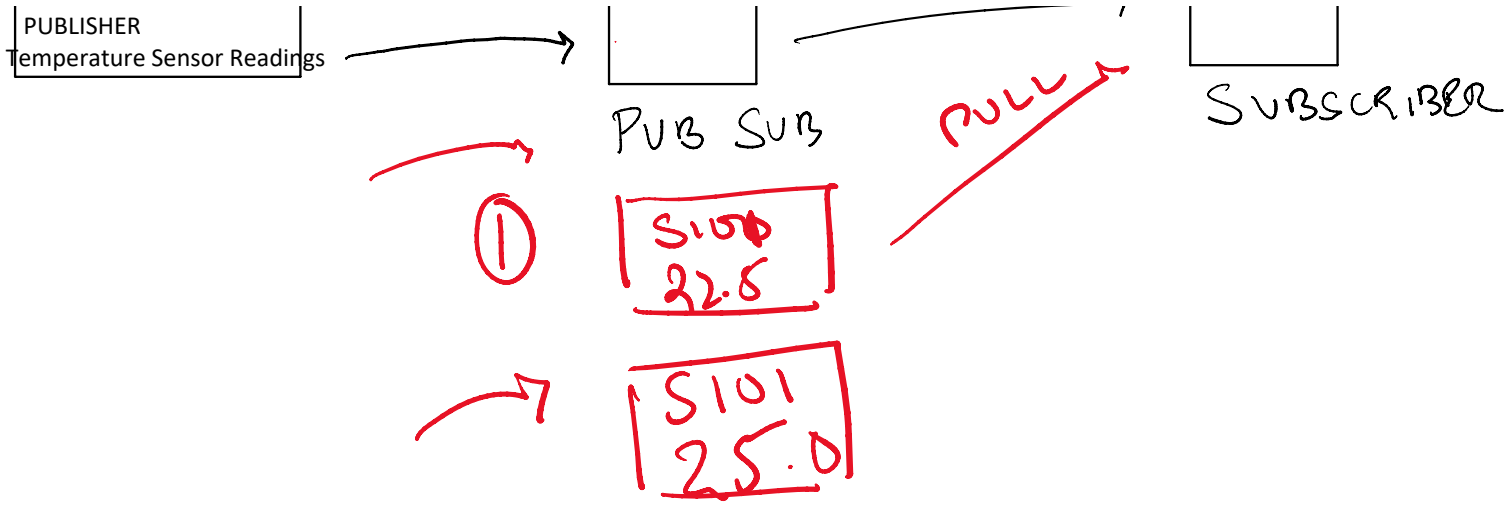
- Asynchronous Messaging Service. PUB = PUBLISHER | SUB= SUBSCRIBER



Core Concept

1. **Topic** - A named resource to which messages are sent by publisher.
2. **Subscription**
3. **Message**:
   a. Pull
   b. Push
4. **Acknowledgement or Ack**





GCP Dataproc

- Hadoop Spark/ Flink/Presto or other big data frameworks.

Key Feature

Feature	**Description**
Fully Managed	
Auto Scaling	
Integration	Big Query, Cloud Storage, PubSub and Vertex AI.
Job	Submit jobs - Spark, Hadoop or Flink

Use Case

1. ETL
2. ML - Train models using Spark MLlib
3. **Real time** analytics
4. Data lake processing

Architecture

1. Master Node - Manage cluster and job scheduling
2. Worker Node - Execute tasks

GCP Dataflow

- Apache Beam
- For streaming and batch data processing

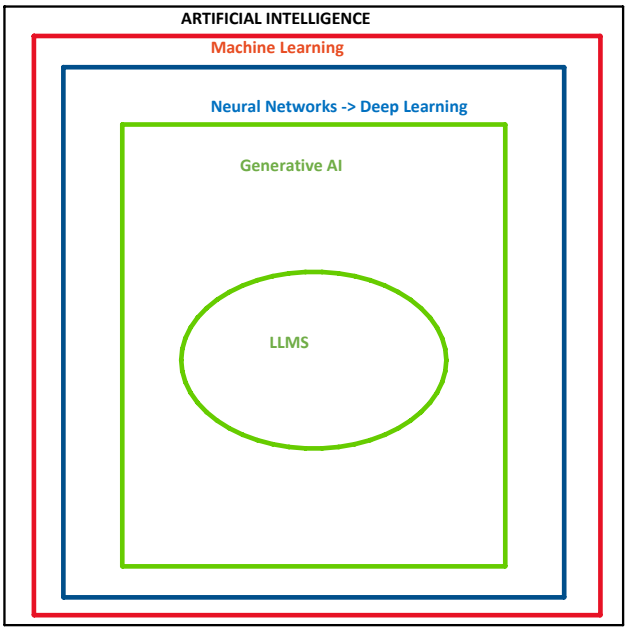
Apache Beam	Apache Spark	Apache Hadoop
Unified programming model for batch and stream data processing.	Fast and general purpose.	Distributed storage and processing framework. MapReduce
DataFlow	Dataproc	Dataproc

Demo:

Step1 - Simulate real time sensor data using PubSub
Step2 - Use dataflow to filter high temperature readings and store them in Cloud Storage
Step3 - Use dataproc [spark] to process and analyze the filtered data and load it to BQ.

Vertex AI

- This is the service that supports ML model development, enabling end user to build and deploy models.



Gen AI - Creates new content from what it has learned from the existing content.
Gemini - Is a multimodal AI tool facilitated to data acquisition and analysis.

ML Phases

1. ETL data into BQ.
2. Select and preprocess the data.
3. Create a model
4. Evaluate the performance of the model.
5. Use the model in production