



T7: PubSub

PubSub

Servicio de mensajería q permite la \leftrightarrow asincrona entre quienes producen el P (publisher) \leftrightarrow reciben (subscriber)



TOPIC: tema \rightarrow "feed" sobre una colección de P
↳ organizan & categorizan los P para suscriptores

Mensajes \rightarrow se compone \hookrightarrow datos

SCHEMA (optional) \rightarrow def. formato de info. q se \leftrightarrow

A \rightarrow 4 tipos

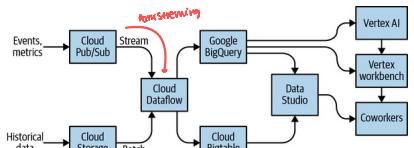
SUSCRIPTORES

PULL \rightarrow **que me devuelves?** \rightarrow reader P de forma activa (releaciones: offset, cuando recibo, q no soy tiempo, control sobre consumo info., q intento q no es lento)

PUSH \rightarrow **yo te mando cuando quería?** \rightarrow publishes P de forma activa real-time o process. datos

BigQuery \rightarrow tipo de suscripción push donde se envía a GCP

EJEMPLO DE USO:



Ciclo de vida de Machine Learning

Ingestión de Datos

Cloud (ex)
GCS
Pub/Sub
API

Comprender SIEMPRE para la ingesta de datos?
Si, hacer un pipeline de ETL

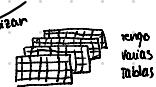
Transformar

BIGQUERY
Cloud (ex)
Composer (con Activadores)

↓
datos

EDA: Análisis, Exploración de datos

Numpy / Pandas
Spark DF
H2O
TensorFlow / Matplotlib
BigQuery



Preparación Tablas

partition train/test

BigQuery
G Cloud, Composer

Objetivo: Ejecutar recurrente manejo (ej: todos los meses)
Para tener las tablas por eso comprobamos

me quiero guardar en una
para hacer el train/test
Left-Join, Subselect, ...

Machine Learning

Ingestión

H2O / Pandas / Sklearn

Preprocesamiento

Feature Engineers / outliers
selección

Sklearn (OneHotEncoder(), Standardizer, ...)
SparkML (Vectorize, OneHot, ...)

Algoritmo

- sklearn, H2O, PyTorch, SparkMLlib
- Modelo \rightarrow objeto \rightarrow Guardar

Producción

Algoritmos train

1 Ingestión de Datos

2 Procesar los datos:
↳ Normalizar: $\frac{x - \bar{x}}{s}$
↳ lo usamos q he hecho para entrenar (using sin select, sin variables, los pocos para directamente)

3 Objeto Modelo + Predecir

↳ Resultado \rightarrow BigQuery / GCP

Evaluación

Ejemplo: Algo q: Prob cada mes q un A me haga un impago

Mes 2 \rightarrow coges datos mes 2 \rightarrow 3/10/xxx - Impago
coges string mes 1 \rightarrow 3/0/xxx - Impago

lo comparas

cliente | predicción (mes 1) | impago (mes 2)

4 0.15
2 0.21
3 0.2
4 0.3

Métricas: precision, recall, f1-score, AUC, ...

¿dónde puedo hacer?

↳ Jupyter Notebook estás + tengo otras entidades (ej: segmentación)

↳ Analizarlo para recomendar \rightarrow q ha salido ese resultado?

↳ EDA: sobre files 1, ... Analizar q ha pasado ese mes

↳ Identificar errores, ... Anotar el problema, ... Corregir el cliente, ... Problema de gestión