Data Analytics Pipeline

from raw data to insight at scale

LUCa Casali & Carlo Zamagni

Chi Siamo & Cosa facciamo

Ing. Casali Luca

Head of Cloud Platform, Ai & Data/Analytics *ho la responsabilità di gestione ed evoluzione della piattaforma cloud, bigdata, ai and analytics.*

@lc2978

Ing. Carlo Zamagni

Senior Software Architect - Cloud Platform Lead Architect mi occupo di studiare, definire ed evolvere l'archiettura della piattaforma Cloud di Technogym

@czamagni

Chi è Technogym: The Wellness Company

Da un garage a una compagnia internazionale

1986

2022

55 000 installazioni **nel mondo**, in 35 000 centri benessere e 20 000 abitazioni private. L'azienda impiega 2000 collaboratori in 14 **filiali** in Europa, Stati Uniti, Asia, Medio Oriente, Australia e Sud America ed esporta il 90% **del** fatturato in 100 paesi.

Technogym Overview

TECHNOGYM®

The Wellness Company™

https://www.technogym.com/

Technogym Village, the new garage for the new era.

Nerio and Pierluigi Alessandri

Chi è Technogym: The Wellness Company

Technogym è riconosciuta in tutto il mondo come azienda leader nella **fornitura di tecnologie, servizi e prodotti di design** per il settore Fitness e Wellness, grazie all'offerta di una gamma completa **di attrezzi** per l'allenamento cardio, forza e funzionale, servizi (post vendita, formazione e consulenza, interior design, marketing support e finanziamenti) **oltre ad una piattaforma digitale cloud** che consente agli utenti di connettersi alla propria personale esperienza Wellness in qualunque luogo siano, tramite i prodotti Technogym stessi oppure con dispositivi mobile.

Technogym: Fornitore di servizi digitali

- Forniamo servizi digitali ai nostri clienti con professional apps, su smartphone, web e tablet.
- + 450K equipment connessi nel mondo che comunicano con la piattaforma mywellness cloud per monitorarne il funzionamento
- 24 milioni di utenti registrati di cui 2.2 milioni di utenti attivi mese
- Esperienza di allenamento totalmente personalizzata completamente integrata in ogni location nel mondo (netflix video)
- **Technogym ha una area di sviluppo e gestione sw che è diventata uno fulcro del dipartimento r&d della azienda**

TECHNOGYM GLOBAL POSITIONING

FACTS & FIGURES

14 **branches**
with a staff of 2,300 employees

55 million
users train with Technogym

80 **distributors**
in more than 100 **countries**

22 million
registered users

485,000 installations
400,000 private & 85,000 professional

Over 38 years of **history**

3 million pieces of
equipment installed

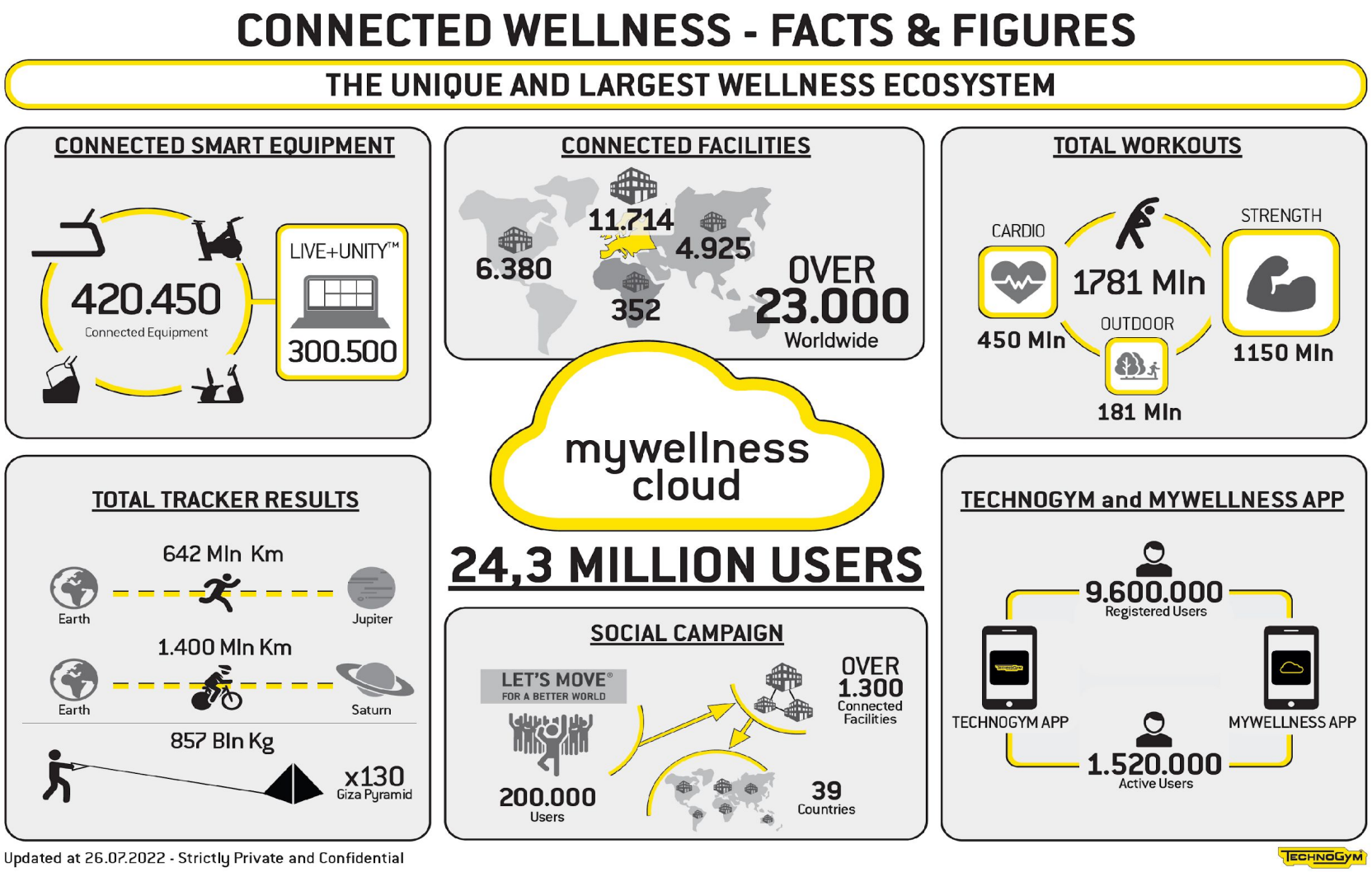
Official supplier to
8 Olympic Games

Champions train with Technogym

8 times partner of the Olympic Games

Partner of top teams and sport events

The largest Wellness Community in the world, constantly growing



Cosa vi presentiamo oggi?

Una pipeline di ***Data Analytics Pipeline at scale*** realizzata attraverso l'utilizzo della piattaforma Google Cloud

Data Is Essential To Digital Transformation

Data is a key pillar for digital transformation because every interaction in the digital world generates data

source: Forbes 2022 Saket Sharma

Obiettivi

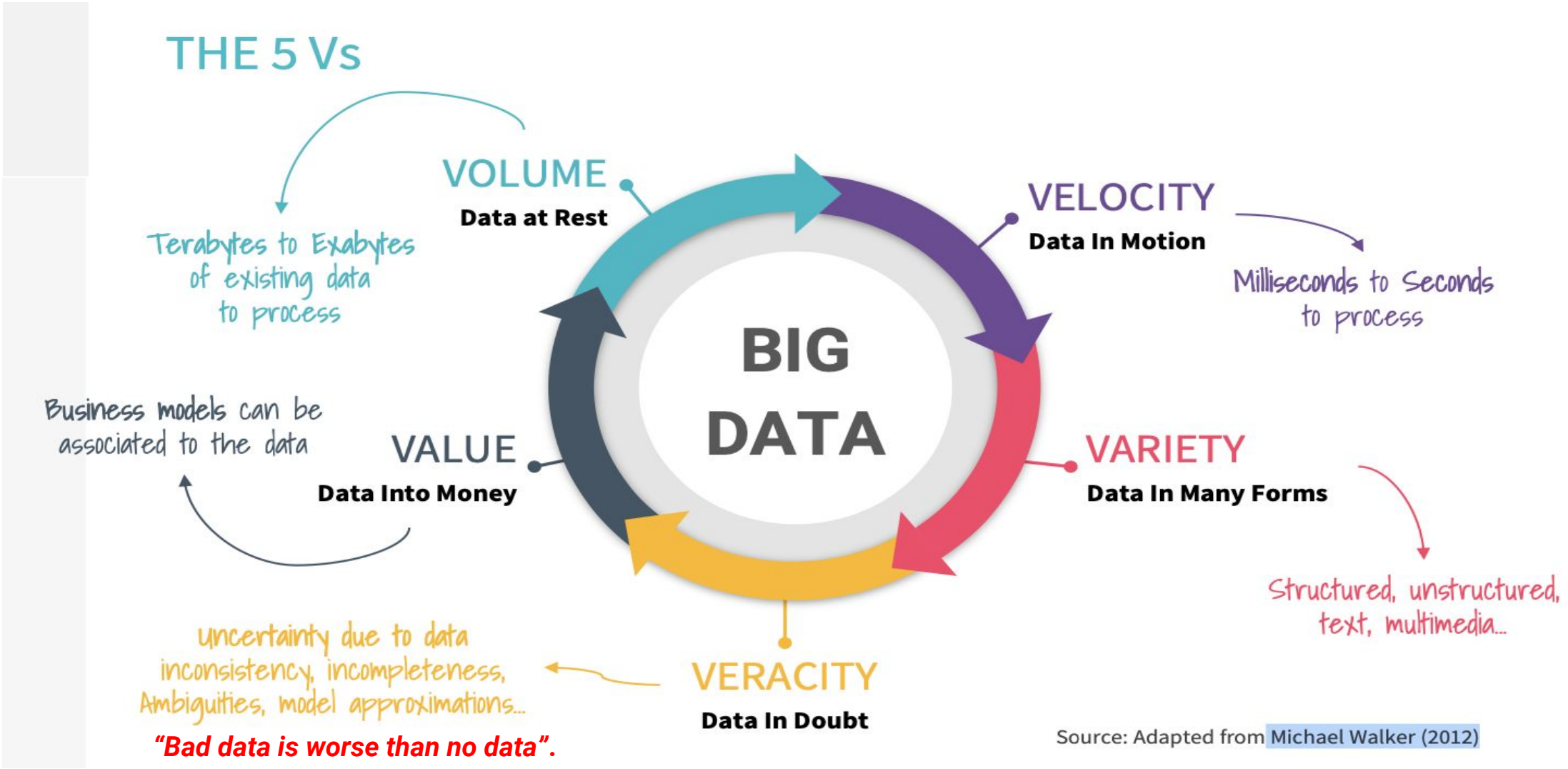
- Definire cosa è una “data analytics pipeline”
- Capire come si struttura una “data analytics pipeline” at scale
- Esplorare le potenzialità di una piattaforma PASS (platform as a service)
  - Costi vs Benefici vs Tech Specs
- Non solo teoria ma un real case da realizzare “insieme”
- *Interagire attivamente con i relatori!!!!*

What is BigData?

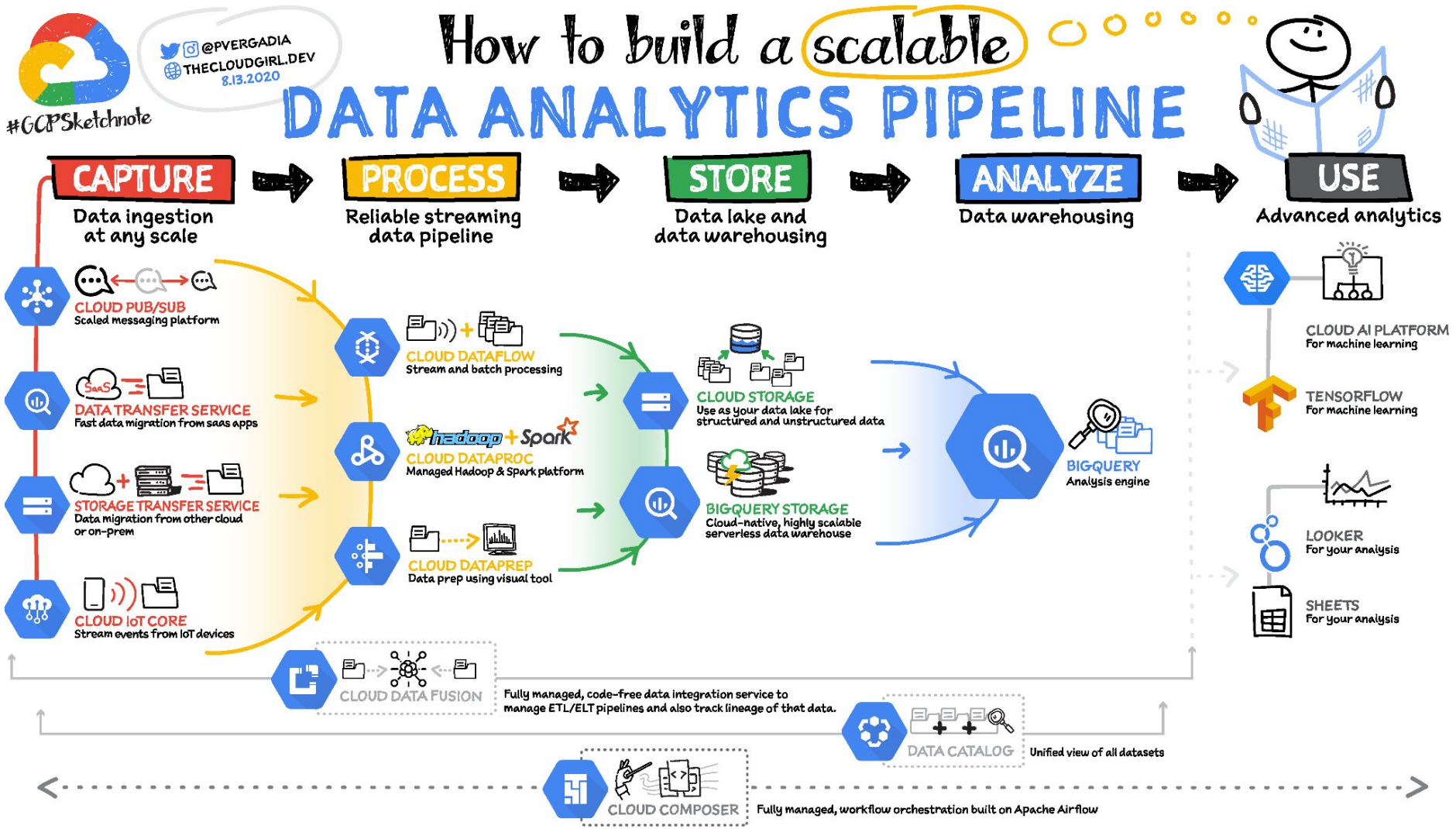
High-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Source: Gartner IT Glossary

What is BigData?: Le 5 V



The Big Picture



Internet of Things - a definition

A distributed system in which objects in the physical world are connected to the Internet.

From remote meters that could be monitored via telephone lines in the ‘70s,

to M2M industrial solutions developed in the ‘90s,

to the endless “internet enabled” devices available today.

From proprietary solutions or industry-specific protocols

to IP-based networks and Internet Standards

Internet of Things - enabling factors

Ubiquitous, low-cost Internet connectivity with high speed and bandwidth

- good luck trying to monitor your 4K security camera via a dial-tone modem from the ‘90s :)
- real time or near-real time information and actions

Globally adopted IP-based networking

- internet connectivity via standard protocols: Ethernet, WiFi, BTLE, …
- devices from different brands can autonomously create mesh/ad-hoc networks (LoRaWAN, Z-Wave, Zigbee, …)

Cheap computing power combined with miniaturization

- a 35$, credit card sized, RaspberryPi board can outpower a desktop computer from 10 years ago
- 1$ micro controllers are powerful enough to connect via WiFi and cypher the connection using SSL (*see Demo Time*)

Internet of Things - enabling factors

Evolution of databases and data analysis techniques

- every year the humanity produces more and more data:

> *“ The amount of data we produce every day is truly mind-boggling. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT).* ***Over the last two years alone 90 percent of the data in the world was generated****.”*
>
> (Forbes - http://bit.ly/3AUNYAc - 2018)

Rise of Cloud Computing

- cheap, remote, on demand, computing power
- SaaS
- PaaS
- (*see Demo Time*)

Internet of Things - scenarios

- **Industry/Logistics 4.0:** joined industrial automation with big data and artificial intelligence unlocking themes like “predictive maintenance”, “smart factories”, and so on.

- **Domotics:** a building can be remotely controlled and sensors and actuators are able to take autonomous “*smart*” decisions based on the data they exchange.

- **Healthcare:** telemedicine and medical robots. Personal sensors, medical certified or consumer grade (just think at your Apple Watch or FitBit).

Internet of Things - scenarios

- **Retail:** goods can be tracked in real time, inventories are automatically updated; pilot projects of unsupervised stores are already in place in some cities.

- **Smart Cities:** data collected from mobile devices, sensors and any sort of equipment are used to monitor, manage and improve operations on urban assets and resources in a “smart” way.

Internet of Things - scenarios

- **And of course Fitness and Wellness:**

RAW Data Types and where to find them

Status Data: most devices can “speak about itself”, moreover they report their current status

- “the sound diffusion in your house is currently playing `<insert song title>`”
- “the treadmill nr.4 in the gym has been in use since 20 minutes”
- the wearing status of a component

Time Series: are continuous sequences of values

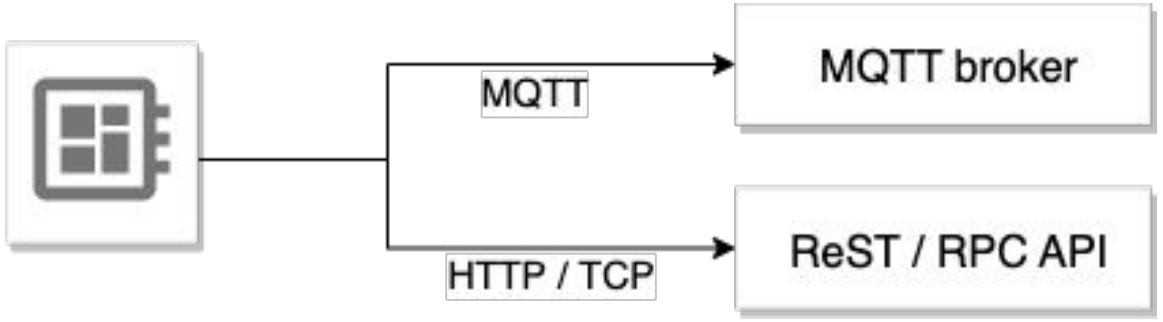
- the temperature in a room
- the continuous monitoring of your heart-rate
- the exchange value of a currency
- the monitoring of industrial devices (nr. of pieces produced, the RPM/power of an engine and so on)

Location Data: commonly found in scenarios like fleet monitoring or fitness tracking

Images/video/files:

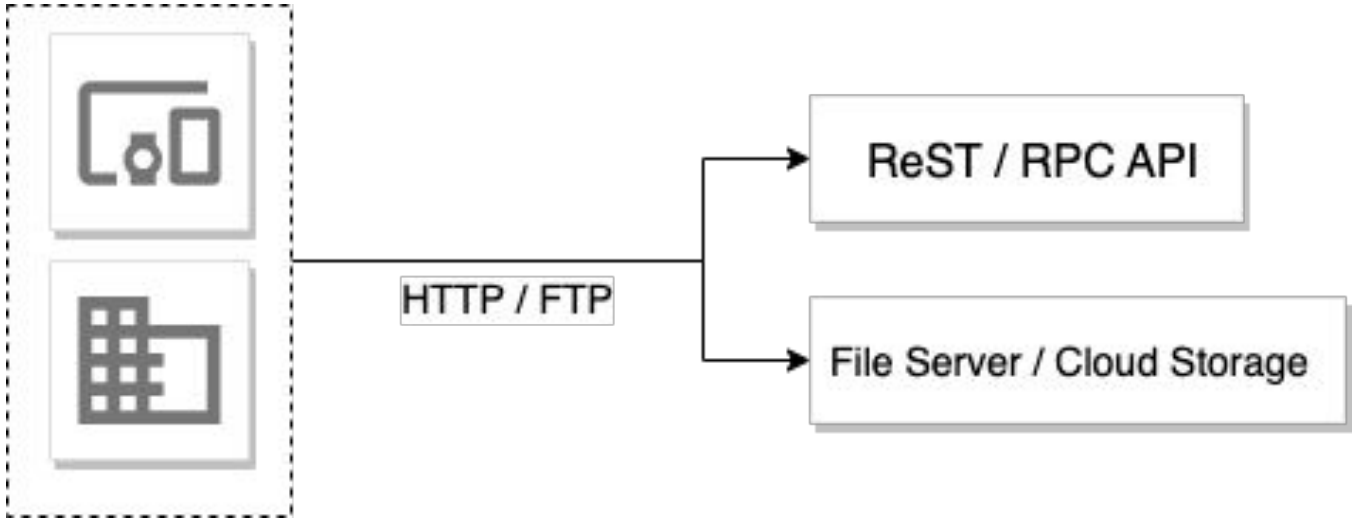
- medical scenarios
- autonomous monitoring

RAW Data - ingestion



- **small data size:** mainly used for commands, single point values, telemetry, time series
- **frequency of communication different from device to device:** from high frequency time series to devices that aren't always connected
- **direction of communication:** in some cases communication needs to be bi-directional

RAW Data - ingestion



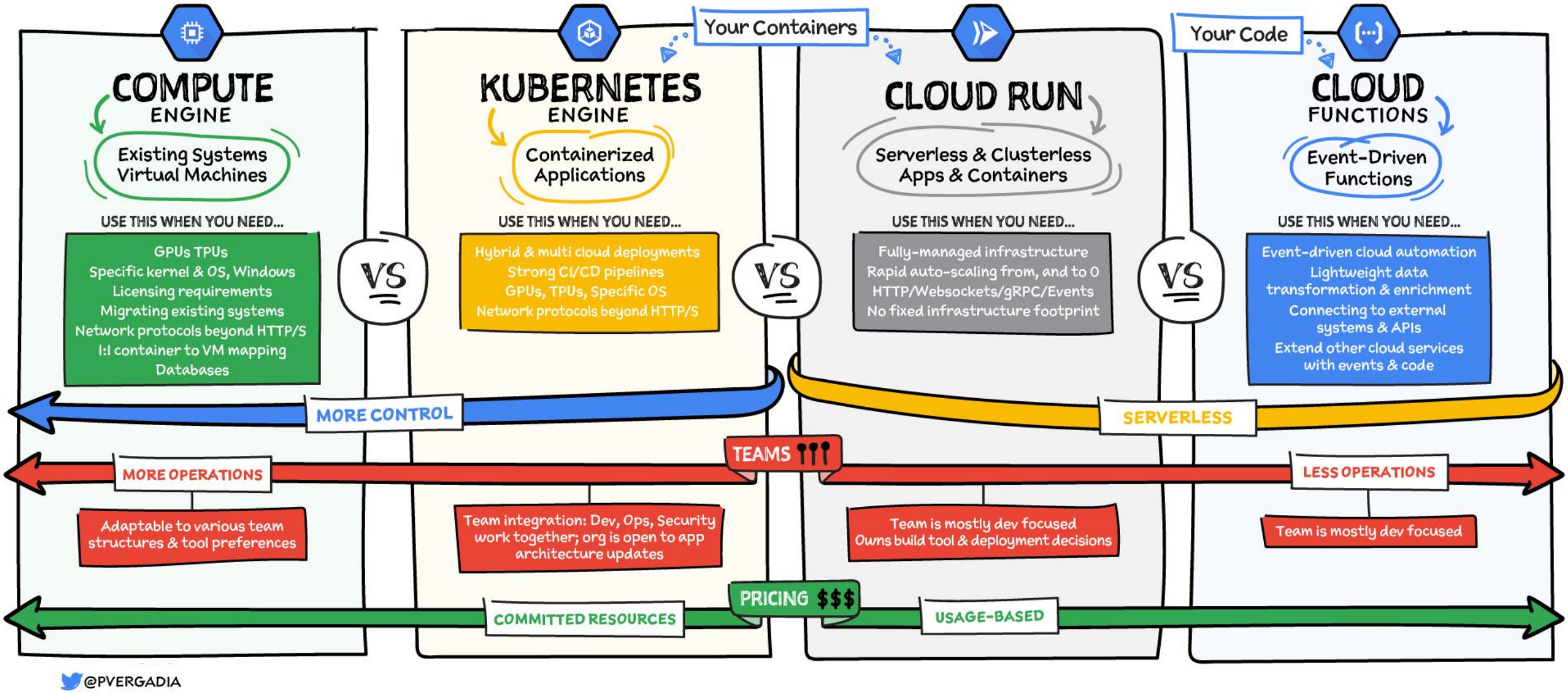
- **structured data:** different representations can be used (JSON, XML, binary files)
- **unpredictable data size:** from small payloads to big files
- **direction of communication:** typically using a client - server approach

RAW Data - ingestion

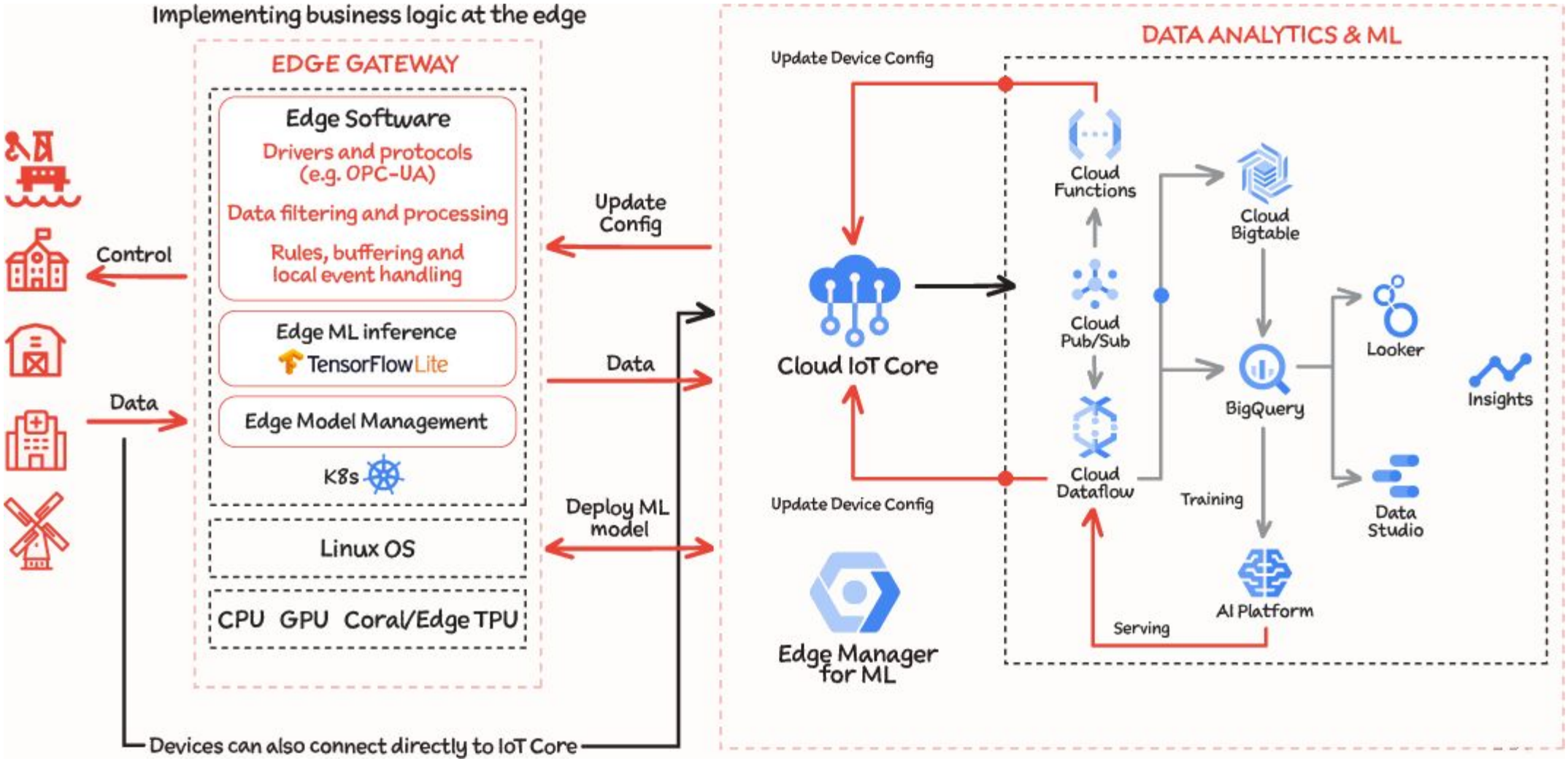
A Cloud Provider offers many solutions to implement this scenario, each one tailored for a specific use case.

Let’s review all the alternatives available on GCP

RAW Data - ingestion on GCP via custom application



RAW Data - ingestion on GCP - Edge devices and MQTT



@PVERGADIA

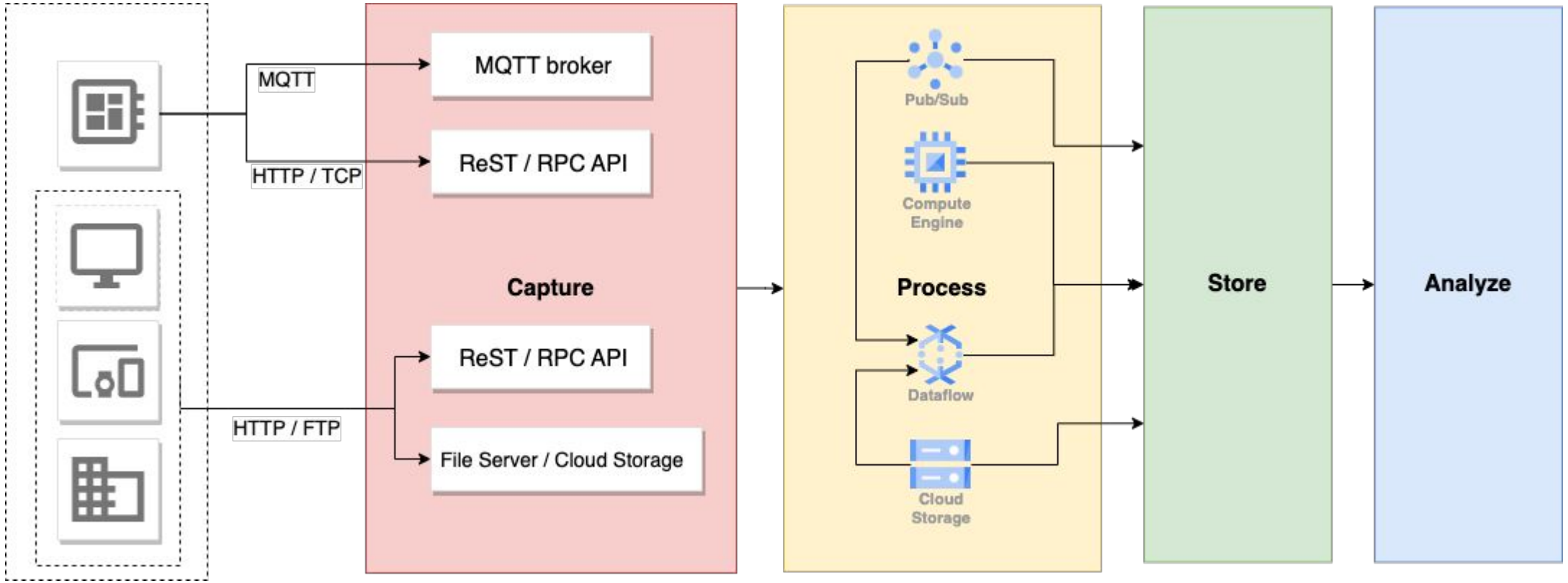
Data Processing

Since now we focused on gathering data from sources and storing them somewhere, now we need need to:

- post process
- aggregate
- validate
- clean
- correlate

The results will be stored into a data lake for further analysis.

Data Processing



Data Processing - some concepts

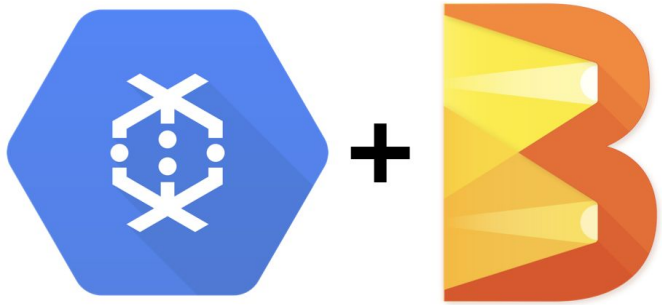
The need for a data processing system in the cloud that is:

- **scalable:** you may have to crunch TB of binary data or billions of database records (e.g. DOR)
- **parallel:** with such data sizes, vertical scaling is not a viable approach
- **streamed:** in many real world scenario data processing need to start as soon as data are ingested
- **portable:** in the cloud environment you should always avoid vendor lock-in whenever possible

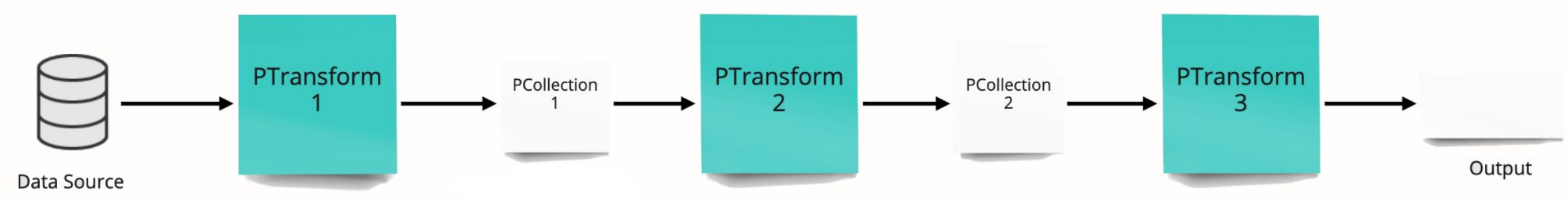
The origins: ***"MapReduce: Simplified Data Processing on Large Clusters"***

Data Processing - Apache Beams and Dataflow

- **scalable:** Horizontally scalable. Autoscaling happens step by step in the middle of a job. As a job needs more resources, it receives more resources automatically.
- **parallel:** can handle batch and streaming data with no need to write different logics separately; it combines them.
- **streamed:** as long as there are no explicit dependencies each step is executed without waiting for the previous one to complete
- **portable:** strict separation of concerns between definition and runtime. Pipelines can be run on Apache Spark, GCP Dataflow, Flink; the SDK supports all most diffused languages such as Java, Python, Golang, Javascript/Typescript.



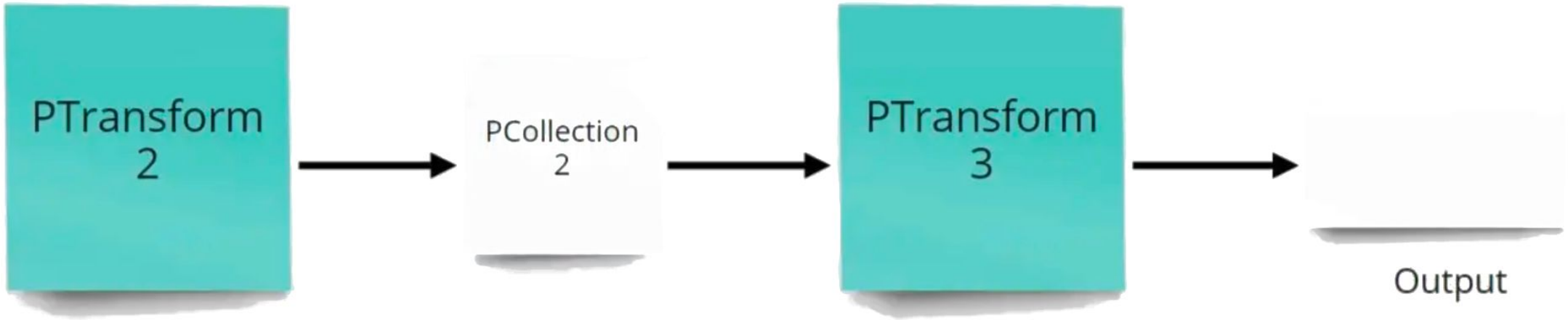
Data Processing - Apache Beams and Dataflow



Pipeline: a *directed acyclic graph* of all the data and computations in your data processing task. It includes reading input data, transforming that data, and writing output data.

PCollection: an immutable, unordered, distributed (as usually doesn't fit into a single machine's memory) dataset of homogeneous data that can represent a defined or a streaming dataset.

Data Processing - Apache Beams and Dataflow



PTransform:

- Handles input, transformations, and output of the data.
- Mathematical calculations, data type conversions, and data grouping steps, as well as performing read and write operations.

Core Beam Transforms

ParDo or ParallelDo (Similar to "Map" phase in Map Reduce):

- Intermediate step in a pipeline. Parallel processing operation.
- Runs a user-specified function on each element in a PCollection.
- May receive addtional data from other PCollection by using side input. Example: Joins.
- Side outputs are useful for additional processing paths.
- Needs code in the form of a DoFN object.

Aggregating with **GroupByKey** (Similar to "Shuffle" phase in Map Reduce):

- Process of computing a result from multiple inputs. Processing collections of key/value pairs.
- Good way to aggregate data that has something in common.

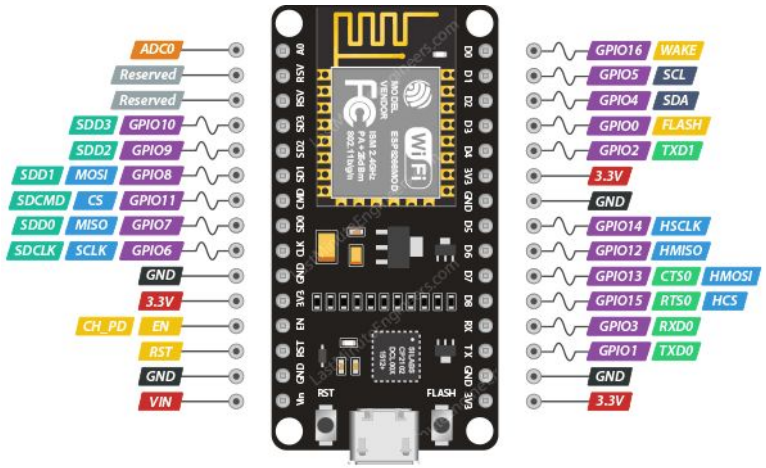
CoGroupByKey

Combine

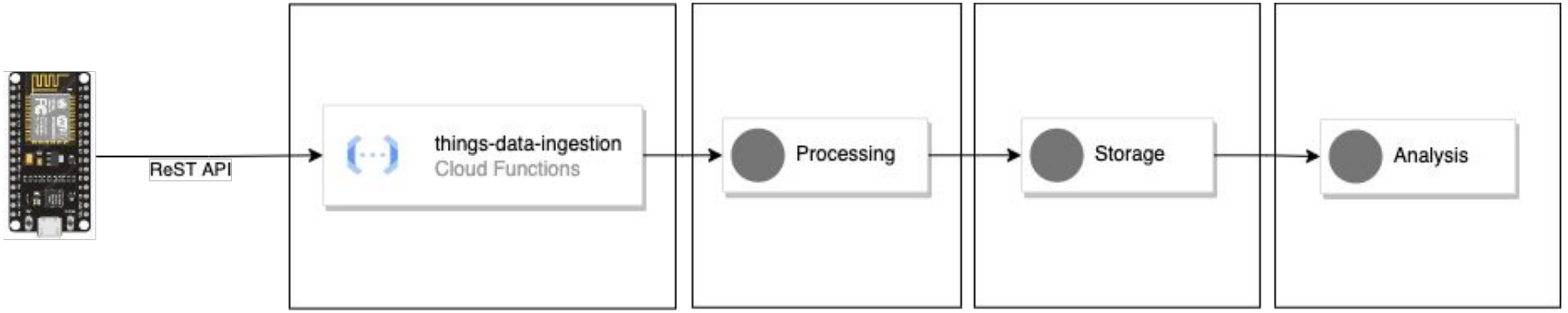
Flatten

Partition

Demo Time - From RAW Data to Insights



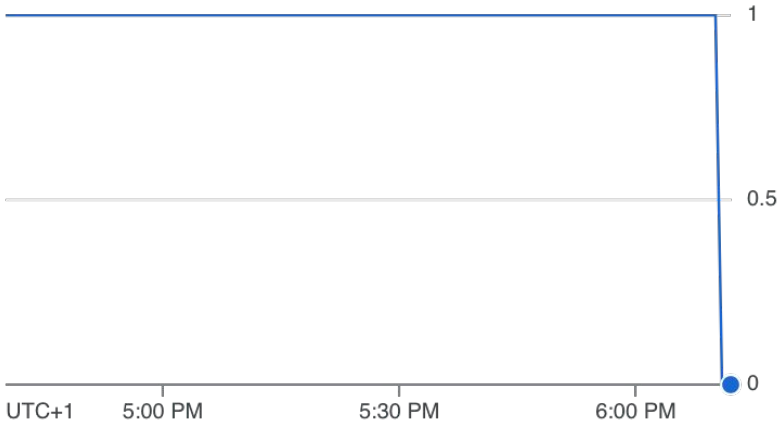
Demo Time - Data ingestion



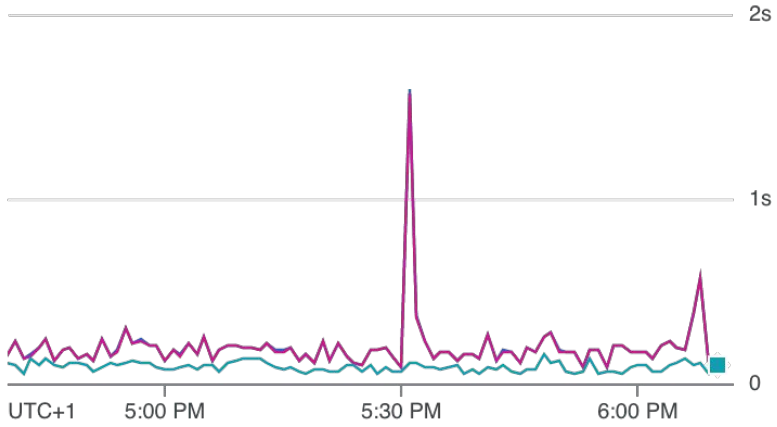

```
const { PubSub } = require('@google-cloud/pubsub')
const pubsub = new PubSub()
const topic = pubsub.topic('projects/technogymplayground/topics/things-data')

exports.entrypoint = (req, res) => {
 topic.publishMessage topic.publishMessage({
   data: Buffer.from(JSON.stringify({data: { message: req.body}}), 'utf8')
 }).then(_ => {
   res.status(200).send(`message: ${req.body.message} value: ${req.body.data}`)
 }).catch(err => {res.status(500).send(err)})
}
```

Demo Time - Data ingestion

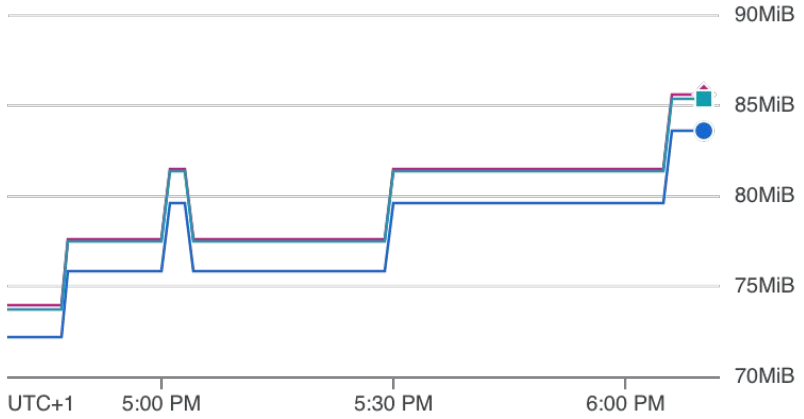
Active instances



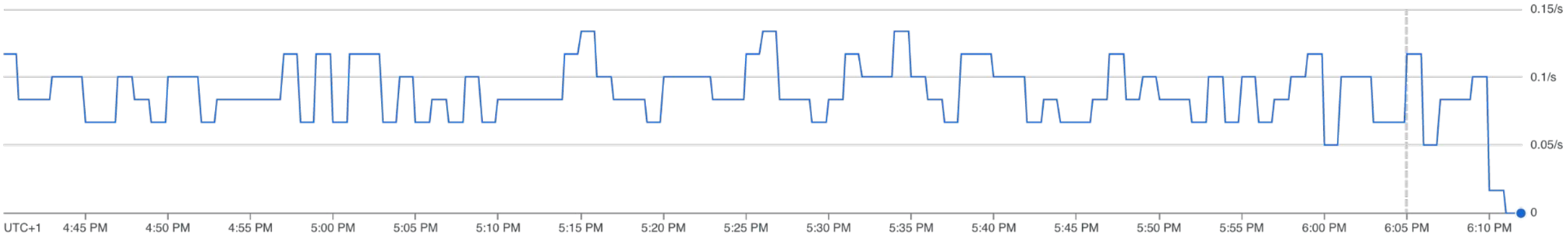
Execution times



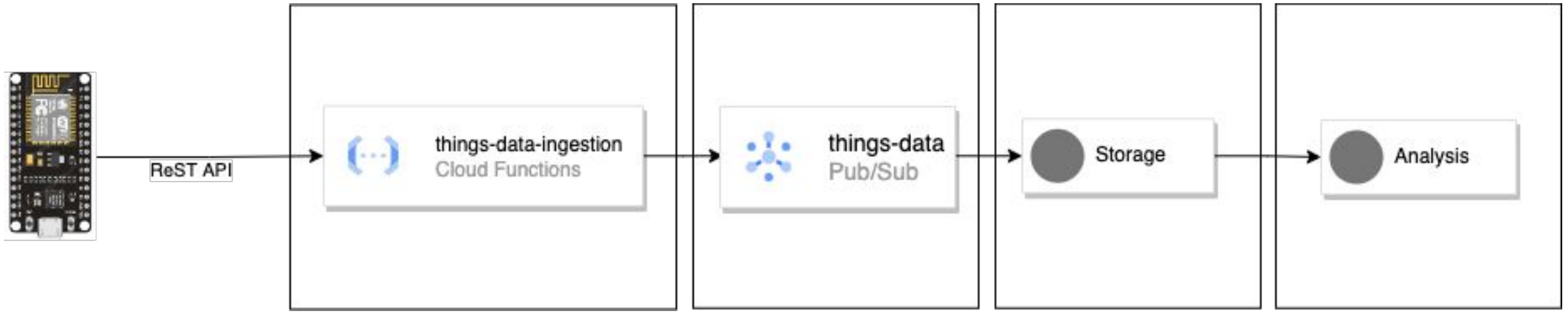
Per call memory

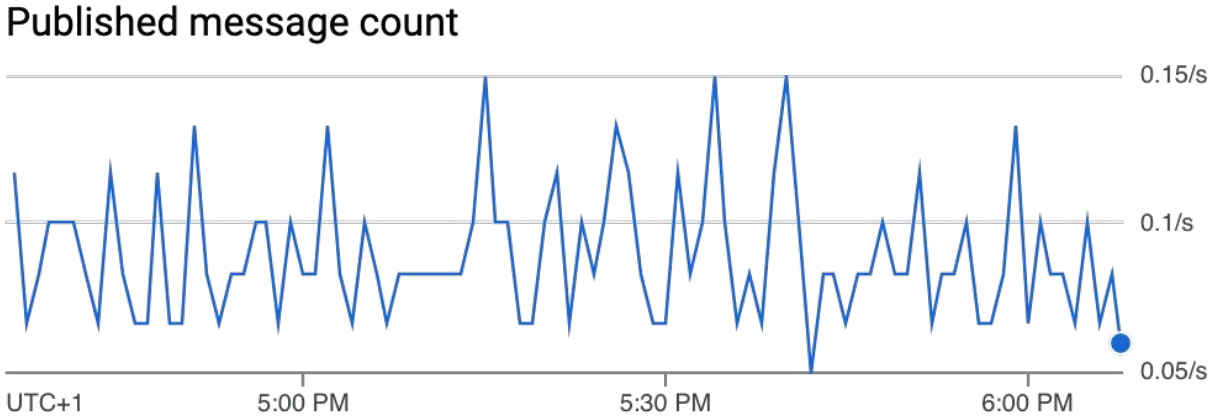


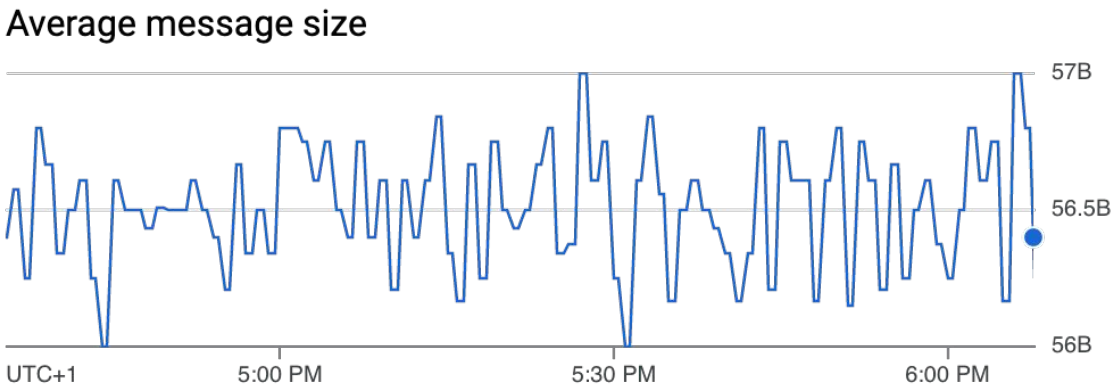
Execution count



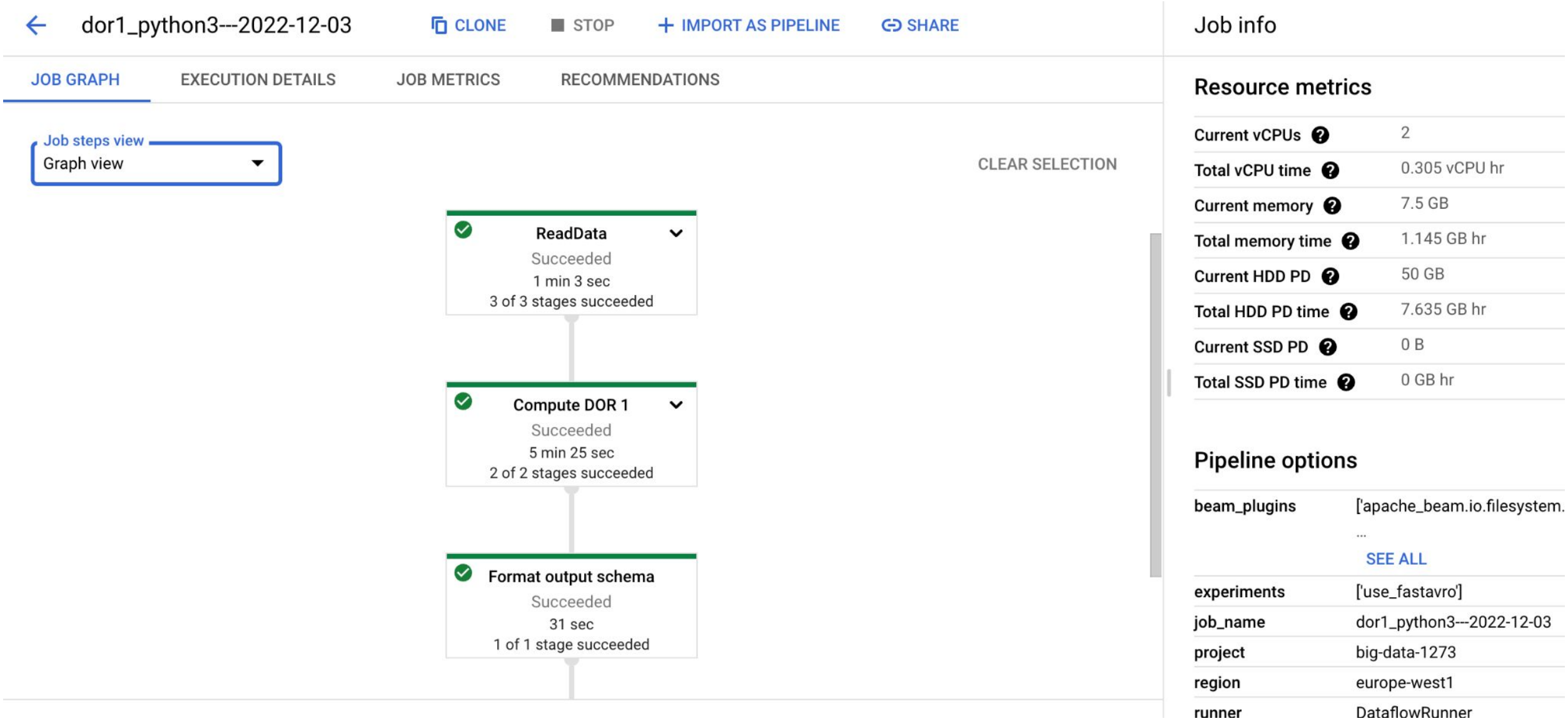
Demo Time - Capture and Process







Demo Time - Capture and Process - an actual enterprise process



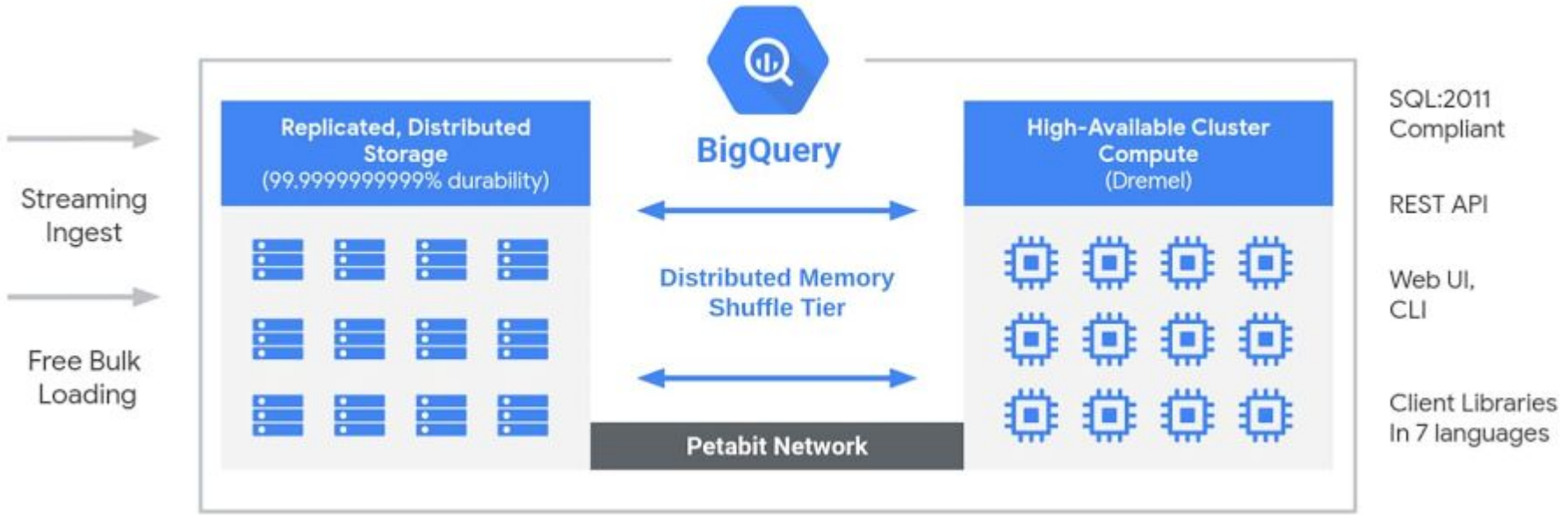
Store: “make data persistent, durable and reliable”

- **Availability**
  Availability can simply be understood as system uptime
  Highly available systems are designed to minimize downtime and avoid loss of service
- **Durability**
  Durability refers to the continued persistence of data. Businesses will have long-term data retention goals. This is achieved by improving durability of the data and the storage infrastructure preserving it.
- **Reliability**
  Reliability is typically associated with the infrastructure storing the data. It refers to the probability that the storage system will work as expected.
- **Fault tolerance**
  Fault tolerance is similar to the concept of availability, but it goes one step further to guarantee zero downtime.
- **Consistency**
  Data consistency, as a support for data integrity, ensures users of the data share the same view of the data, including changes that were made by the user and changes made by others

Google BigQuery: Datalake, Data warehouse

Big Query

- Fully Managed Big Data Analytics Service
- Support SQL
- Fast
- Scalable
- Flexible and Familiar
- Security and Reliability



Store - Process - Transform: Demo Time

Google Big Query

Analyze: Data Analytics

- *Looker Studio*
  Self-service business intelligence with unmatched flexibility for smarter business decisions. (Formerly known as Data Studio)
  https://datastudio.google.com/u/0/

- *Connected Sheets*
  Connected Sheets allows you to analyze petabytes of data directly within Sheets.

- *Colab*
  Colab notebooks are Jupyter notebooks that are hosted by Colab
  https://colab.research.google.com/

Demo Time: Analytics tools

Last but not least: Machine Learning at scale

BigQuery ML ti consente di creare ed eseguire modelli di machine learning in BigQuery utilizzando query SQL standard. Non c'è bisogno di soluzioni che utilizzzano python o java ma utilizzando sql

BigQuery ML aumenta la velocità di sviluppo eliminando la necessità di spostare i dati.

BigQuery ML: Modelli

- Linear regression for forecasting; for example, the sales of an item on a given day. Labels are real-valued (they cannot be +/- infinity or NaN).
- Binary logistic regression for classification; for example, determining whether a customer will make a purchase. Labels must only have two possible values.
- Multiclass logistic regression for classification. These models can be used to predict multiple possible values such as whether an input is "low-value," "medium-value," or "high-value." Labels can have up to 50 unique values. In BigQuery ML, multiclass logistic regression training uses a multinomial classifier with a cross-entropy loss function.
- K-means clustering for data segmentation; for example, identifying customer segments. K-means is an unsupervised learning technique, so model training does not require labels nor split data for training or evaluation.
- Matrix Factorization for creating product recommendation systems. You can create product recommendations using historical customer behavior, transactions, and product ratings and then use those recommendations for personalized customer experiences.
- Time series for performing time-series forecasts. You can use this feature to create millions of time series models and use them for forecasting. The model automatically handles anomalies, seasonality, and holidays.
- Boosted Tree for creating XGBoost based classification and regression models.
- Deep Neural Network (DNN) for creating TensorFlow-based Deep Neural Networks for classification and regression models.
- AutoML Tables to create best-in-class models without feature engineering or model selection. AutoML Tables searches through a variety of model architectures to decide the best model.
- TensorFlow model importing. This feature lets you create BigQuery ML models from previously trained TensorFlow models, then perform prediction in BigQuery ML.
- Autoencoder for creating Tensorflow-based BigQuery ML models with the support of sparse data representations.

Demo Time: BigQuery ML - Actionable Data

Logistic regression model utilizzando solo costrutti SQL

How to choose AWS or GOOGLE ?

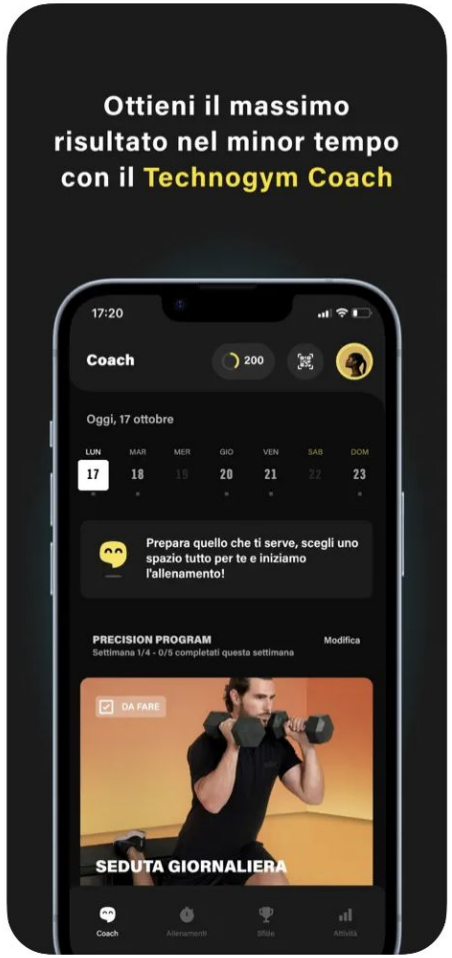
Quali sono i pregi/vantaggi e difetti delle due piattaforme?

La nostra esperienza dice…..

Question and Answer?

- ★ lcasali@technogym.com
- ★ cazamagni@technogym.com

Gift for you



3 mesi di allenamenti con Technogym App

https://technogym.page.link/snxhNUbFXMvniJEDA

SCAN ME

Grazie di aver partecipato

Platform as a Service

Platform as a service (PaaS) or application platform as a service (aPaaS) or platform-based service is a category of cloud computing services that provides a platform allowing customers to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app

Vantaggi e Svantaggi

The advantages of PaaS are primarily that it allows for higher-level programming with dramatically reduced complexity; the overall development of the application can be more effective, as it has built-in/self up-and-down ramping infrastructure resources; and maintenance and enhancement of the application is thus easier.

Disadvantages of various PaaS providers as cited by their users include increased pricing at larger scales,[23] lack of operational features,[24] reduced control,[24] and the difficulties of traffic routing systems.[25]

Software as a service

Software as a Service (SaaS) is a form of cloud computing that delivers **software functionality** over a network, usually the internet.

- The software is running on a server(s) in a data center not physically in the same location as the user - it can be anywhere on the planet
- The user accesses the software using a network connection to the server. Again, the user can be anywhere on the planet and as long as they have a network connection to the server they can use the software's functionality