Exploring the Evolution of Big Data Analytics Technology through Cloud-Native Architectures - A Case Study in Insurtech

葉信和 / Hsin-Ho Yeh
Founder & CEO / Software Engineer @ 信誠金融科技
hsinho.yeh@footprint-ai.com

Tintin
Machine Learning Platform for Everyone

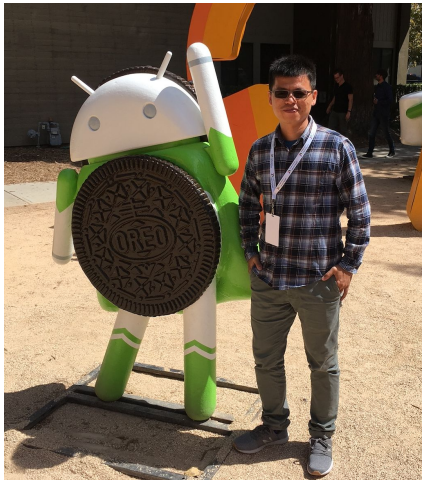
信誠金融科技股份有限公司
XINCHEN FINTECH CO., LTD.

Download Slides

https://bit.ly/47C2YBn

https://github.com/FootprintAI/talks/tree/main/slides

About me



- 2020 - Present at **信誠金融科技**
  - **Tintin**: a machine learning platform for everyone
    - https://get-tintin.footprint-ai.com
- 2016 - 2020 at IglooInsure (16M+ in series A+ 2020)
  - Provide digital insurance for e-conomic world
  - Funded in KUL, Headquartered in Singapore
  - First employee/ Engineering Lead / Regional Head/ Chief Engineer
- 2013 - 2016 at Studio Engineering @ hTC
  - Principal Engineer on Cloud Infrastructure Team
- 2009 - 2012 at IIS @ Academia Sinica
  - Computer vision, pattern recognition, and data mining
- CS@CCU, CS@NCKU alumni

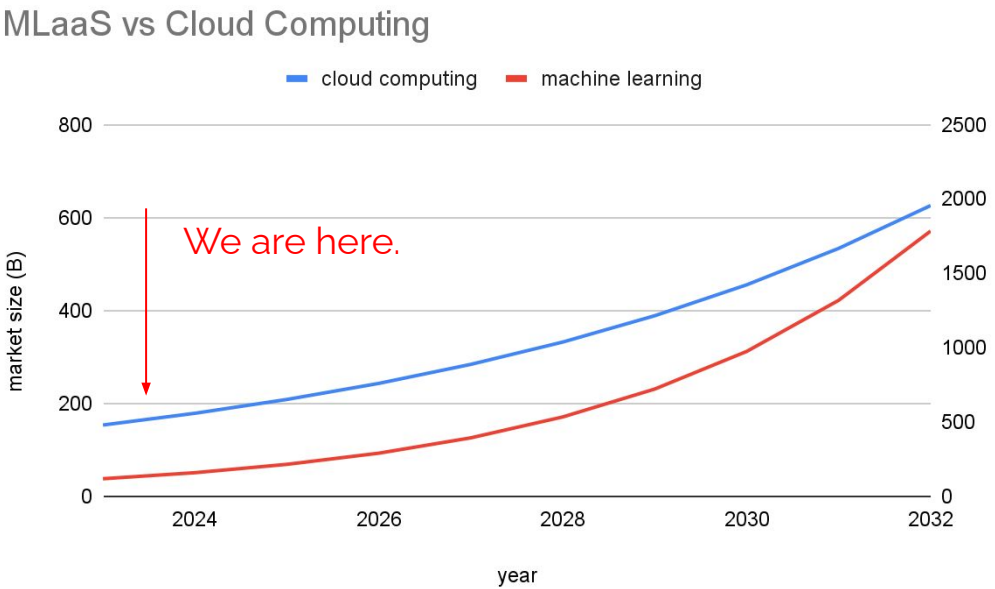
Intro & timeline

- footprint-ai.com (信誠金融科技) is committed to providing Software-as-a-Service for serving AI/ML applications, specializing in:
    - Cloud-native architecture, MLOps, Green Tech, Internet-scale Data Analysis
- Practical experience in taking an idea (zero) to many products (one)
    - Singapore-based insurtech startup iglooinsure (18M USD/2020 Series A, 46M USD/2023 Series B).
- Business partner with IBM Taiwan (IBM.tw)
- Joined Nvidia Inception program in 2021.
- Moved into FinTechSpaces in 2023.

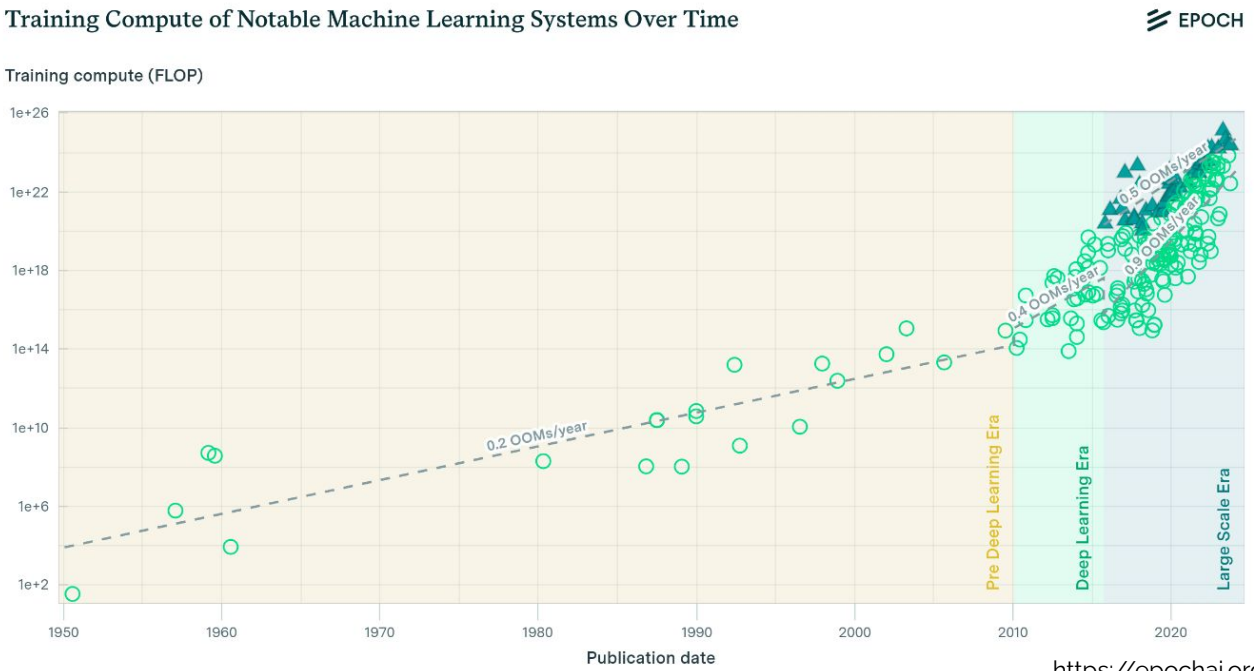
Agenda

- Trends and Observations.
- Data Analytics Sharing from Insurtech Perspective
- The Evolution on Big Data Technologies.
- What is Cloud-native? And how Big Data Technologies is involved?
- What we do at Footprint AI
- Q&A

MLaaS is the future

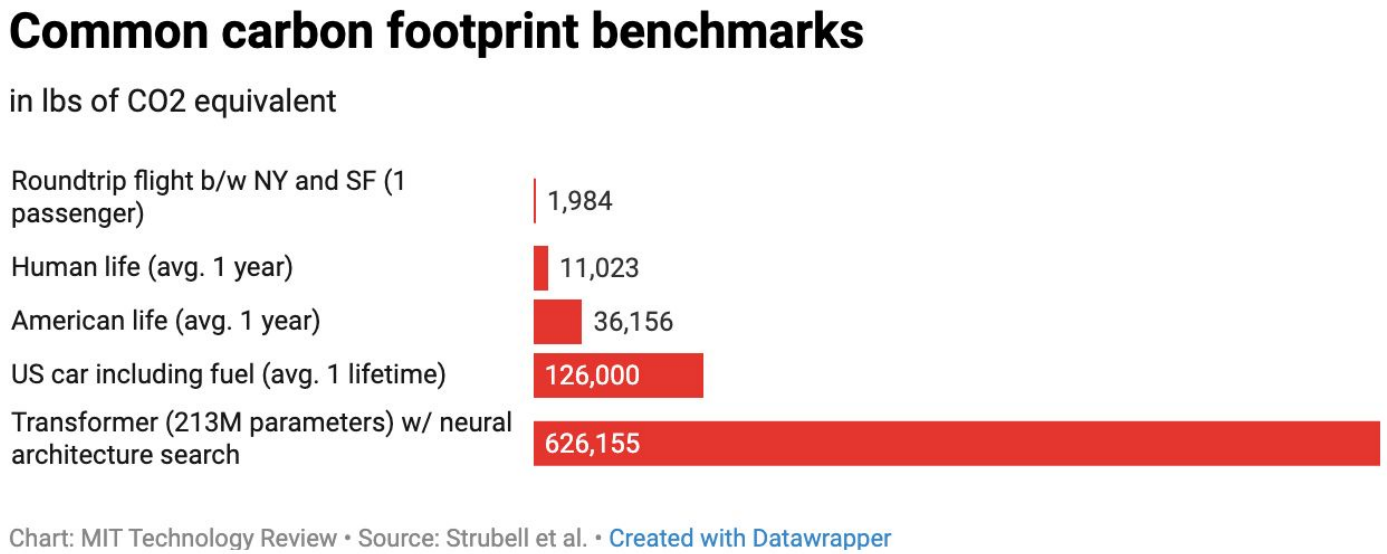


ML Model Growth history



https://epochai.org/blog/compute-trends

But the gas emission is also accelerating …

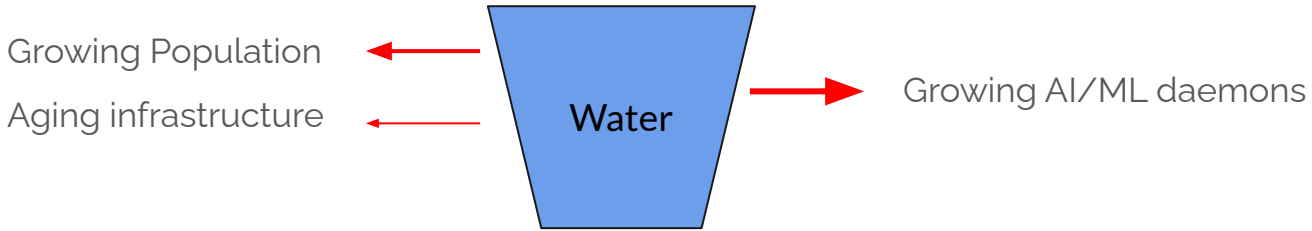


https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

And Water is one of the conflict resources

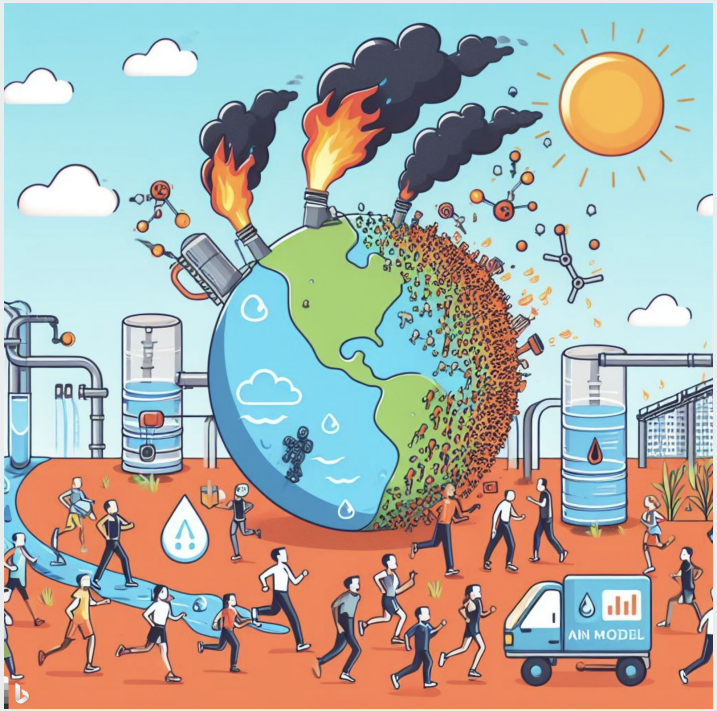
We need AI/ML to accelerate our daily tasks and automation, but …

- Data Centers (DCs) in US are consumed 2% electricity use.
- ChatGPT needs to “drink” a 500ml bottle of water for every simple 20-50 questions and answers (and GPT-4 is even thirstier) [1]



[1] https://arxiv.org/pdf/2304.03271.pdf

What future are you anticipating?



https://www.bing.com/images/create/people-don27t-care-about-carbon-emission-and-water-/654fe2659a0a441e91d57dfd6b950bfa?id=JNCSWj7OssHkQcitT2C9yQ%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay

https://www.bing.com/images/create/create-a-photo-where-an-ai-is-embrace-sustainabili/654fe07a3f8545e099187f40b34fded9?id=Y4n9LElWrhJD%2fMcqQJQxcg%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay

Data Analytics Sharing from insurtech perspective

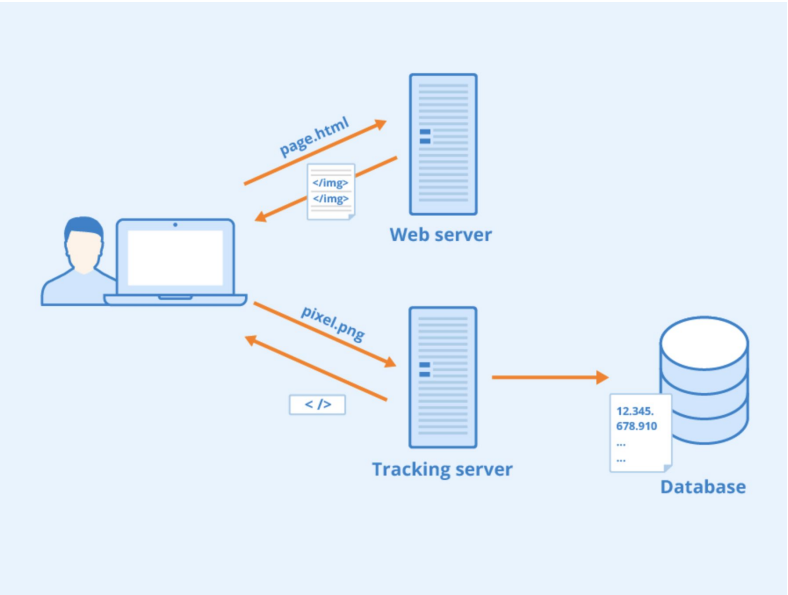
https://www.bing.com/images/create/data-analytics-sharing/1-6565906 54b5d4b658f8bcba58dda1c59?id=K9tlN1ufvyywUmlSj3xqCQ%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay

From insurtech perspective (I)

- Offering a hassle-free protection for user during their shopping experience
  - Cover for the return shipping cost in case the buyer wants to return the product.
- How to predict the user's return rate at the moment when he/she make a purchase?
  - The quality of Shops
    - From shops' previous history
  - The quality of customer
    - From customer's previous history
  - The quality of goods
    - From product's description and photos associated with it.
    - From product's history review.
  - Customer's browsing history before the purchase?
  - Anything else?

Tracking Pixel

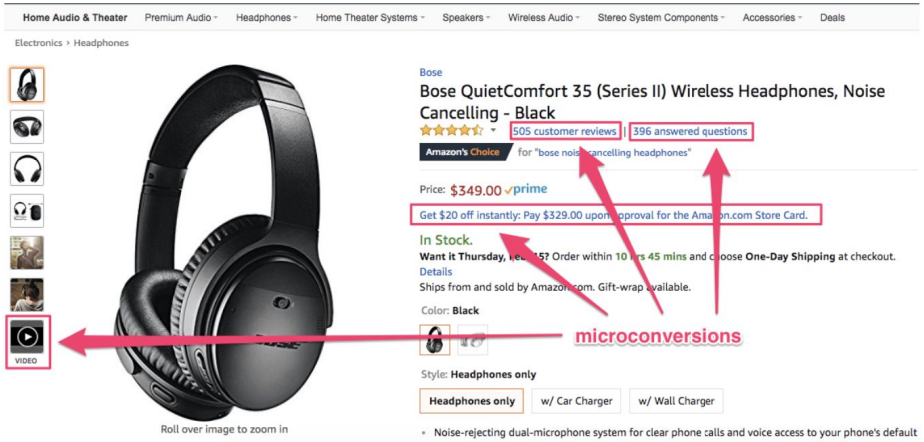
- A tracking pixel is employed to monitor users' browsing activities as they navigate through various pages within the same e-commerce platform.

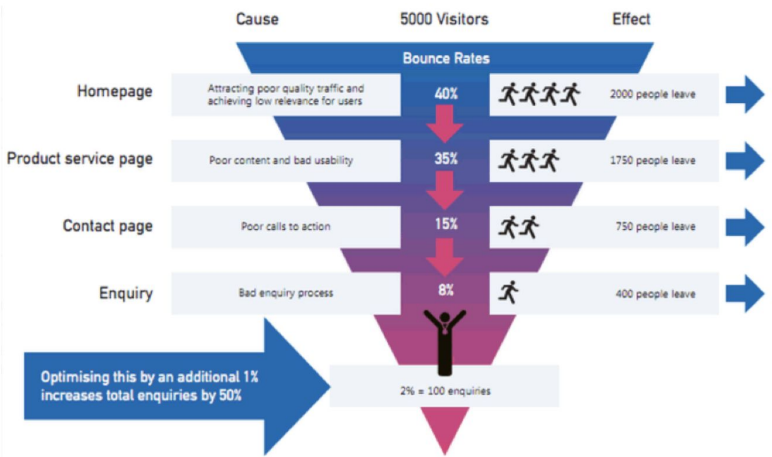


https://www.seobility.net/en/wiki/Tracking_Pixels

Conversion Funnel

- A conversion funnel visualizes the path from potential customers to paying ones

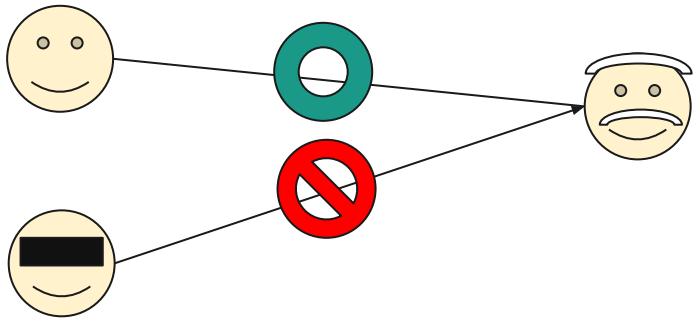




https://www.crazyegg.com/blog/website-conversion-funnel/

From fintech perspective (II)

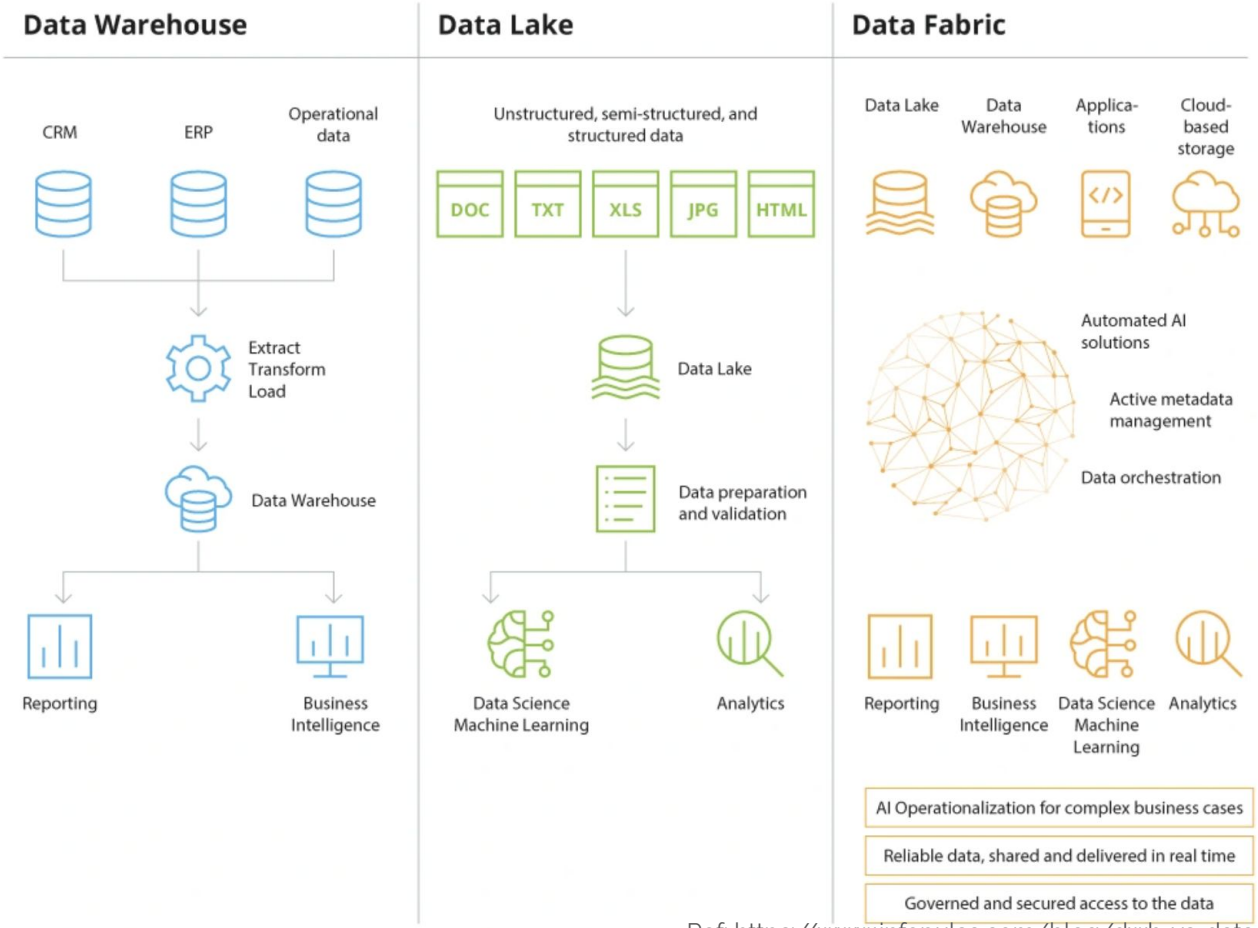
- Providing caller identification to safeguard phone call recipients from harassment or fraudulent activities.
- Applications
  - Financing
    - Establish a proxy for a social score in cases where an individual lacks a valid supporting statement.
    - By analyzing the caller-callee network and activity.
  - User Preference Analysis
    - From pick-or-not-pick activities.



Evolution of big data analytics technology.

https://www.bing.com/images/create/big-data-technologies2c-volume2c-velocity2c-variety/1-65640b8e3e034319883d4d29423e2e7b?id=SopuYuBC7LwjLyDAJaqn3w%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay

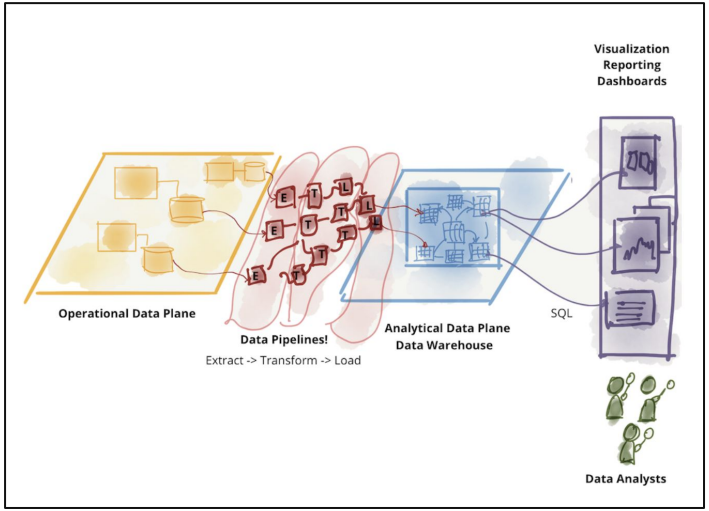
Evolution of big data analytics technology



Ref: https://www.infopulse.com/blog/dwh-vs-data-lake-vs-data-fabric

Data Warehouse

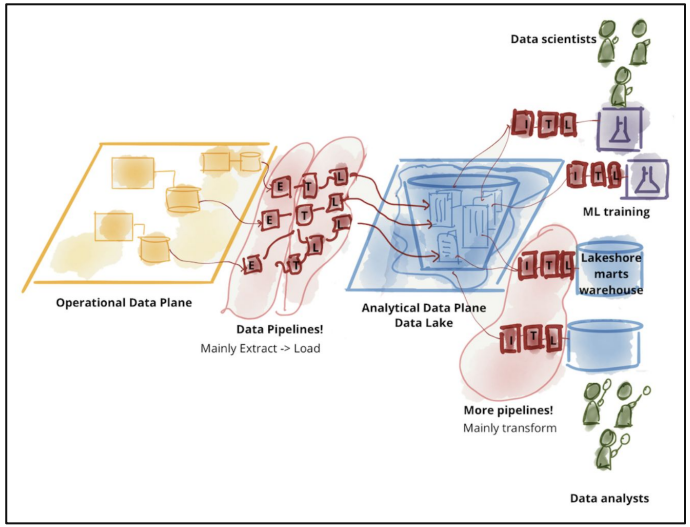
- For **structure data** storage.
  - Bigquery(GCP), Redshift(AWS) and Azure Synapse Analytics.
- Structure Input, ex: transaction data, tracking pixel
- Pipeline-based Process
  - An ETL(extract-transform-load) tool to convert ingested data into the desired format.
    - Browsing trajectory
    - Historical/real-time transactions
- Structure-based query for data consuming, ex: BI report.
  - Risk modeling for a particular shop and customer.



Ref:
https://martinfowler.com/articles/data-mesh-principles.html

Data Lake (I)

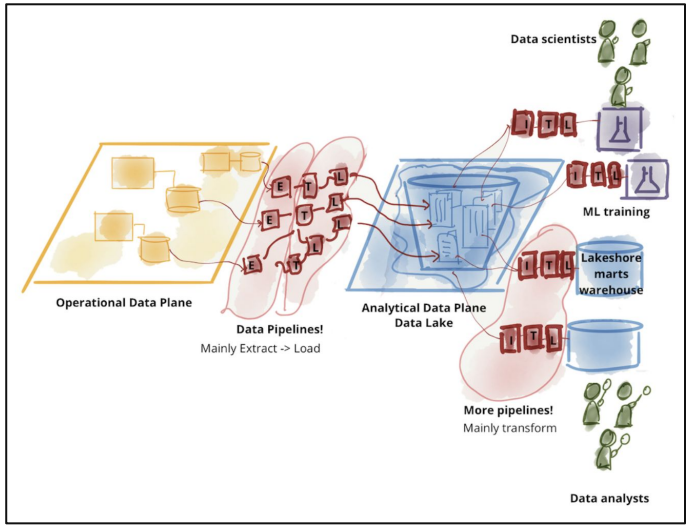
- For **structure, semi-structure, and unstructured** data storage.
  - Spark(Apache), Hadoop(Apache), Dataflow(GCP), Athena(AWS)...
- Caused by low storage cost and leverage open data format (ex. Apache Parquet).
- Drawbacks:
  - Incomplete ACID support. (missing update feature)
  - no enforcement of data governance
  - poor performance optimizations
    - Only rely on indexing, partitioning, and caching.



Ref:
https://martinfowler.com/articles/data-mesh-principles.html

Data Lake (II)

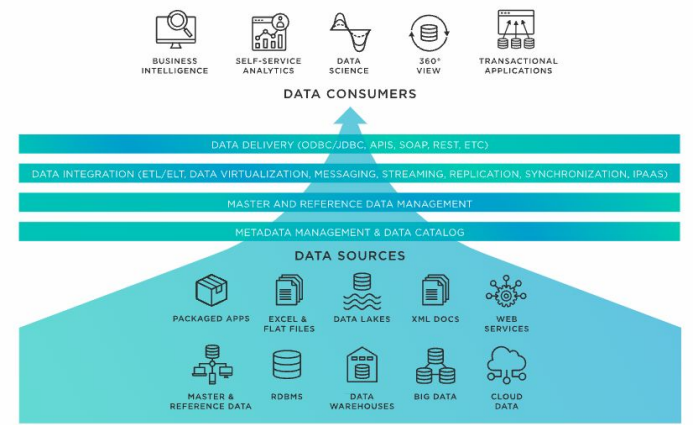
- Modern approach is to hybrid data warehouse and data lake by
  - Use ETL Job to convert unstructured data from data lake to data warehouse for later analysis.
  - It requires engineering workloads and also each step introduced risks and bugs to reduce data quality.



Ref: https://martinfowler.com/articles/data-mesh-principles.html

Data Fabric

- Data governance and compliance
  - Right people access the right data by automatic policy enforcement and regularization compliance.
- Data integration
  - Integrate data from disparate sources into a single format.
  - Allow faster and more efficient access to the data.
    - How to move data from different domains (e.g. from storage to analytics) without actual copy?
- For example: a healthcare organization can track all instances of patient data across its systems and also ensure compliance with regulations such as HIPAA.



https://www.tibco.com/reference-center/what-is-data-fabric
https://mia-platform.eu/blog/data-mesh-vs-data-fabric/

What is Cloud-native infrastructure?

Cloud native technologies empower organizations to build and run scalable applications in modern, dynamic environments such as public, private, and hybrid clouds. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs exemplify this approach.

CLOUD NATIVE
COMPUTING FOUNDATION

Cloud Native Pillars

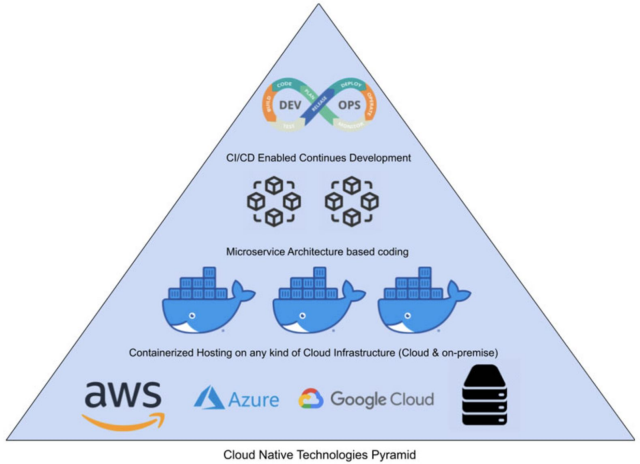
A cloud-native architecture handles containerized workloads and microservices, enabling high scalability and fault tolerance.

Key Components include

- Container and its orchestration
- Serverless architecture
- Continuous integration and Continuous Deployment (CI/CD)

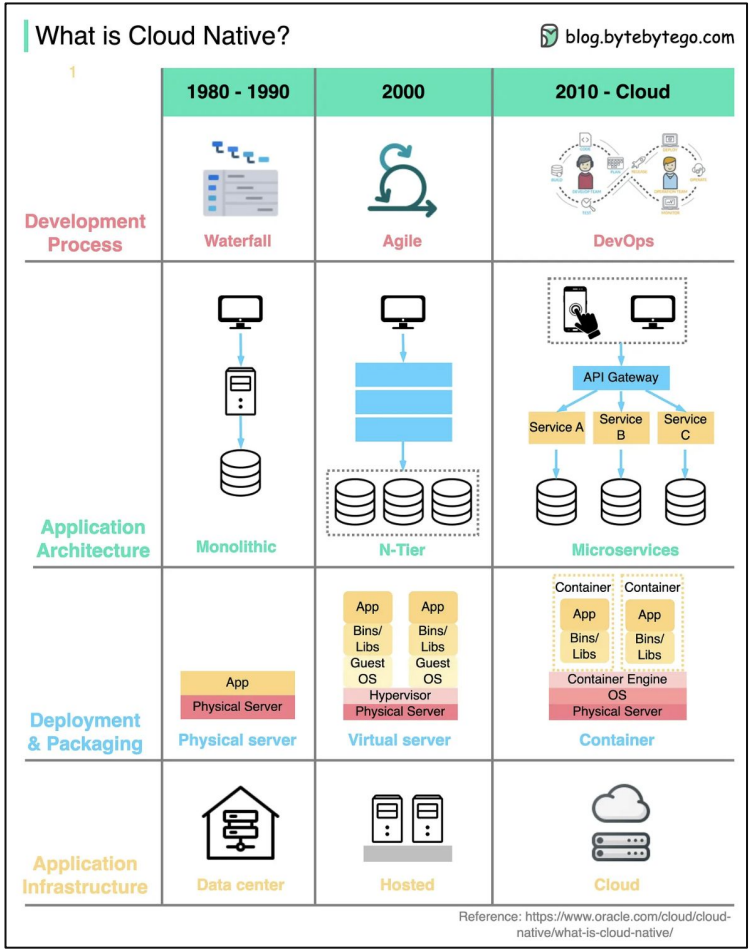
Key Benefits:

- Scalability
- Portability
- Cost-effective



Ref: https://www.influxdata.com/blog/introduction-cloud-native/

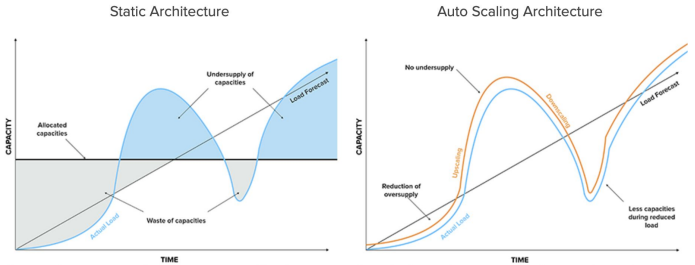
Evolution of cloud native



Ref: https://blog.bytebytego.com/p/ep-39-accounting-101-in-payment-systems

Why Cloud-native play an important role to big data analytics technology?

- Integrating cloud-native and big data technology has become increasingly common in modern data-driven environments, including
  - Scalability: Big data workloads can vary significantly, and the ability to dynamically allocate resources in the cloud ensures optimal performance without overprovisioning.
  - Flexibility and Agility: Containers and container orchestration platforms, such as Kubernetes, allow for consistent deployment across different environments, making it easier to manage and move big data workloads across on-premises and cloud environments.



Static architectures are based on estimated load expectancy and are not flexible enough to adapt to unexpected load peaks or lulls.

Auto Scaling is the most cost-efficient solution for a fluctuating load. High performance and thereby user satisfaction are retained at all times.

Ref: https://www.mgt-commerce.com/cms/aws-auto-scaling-pricing-plans

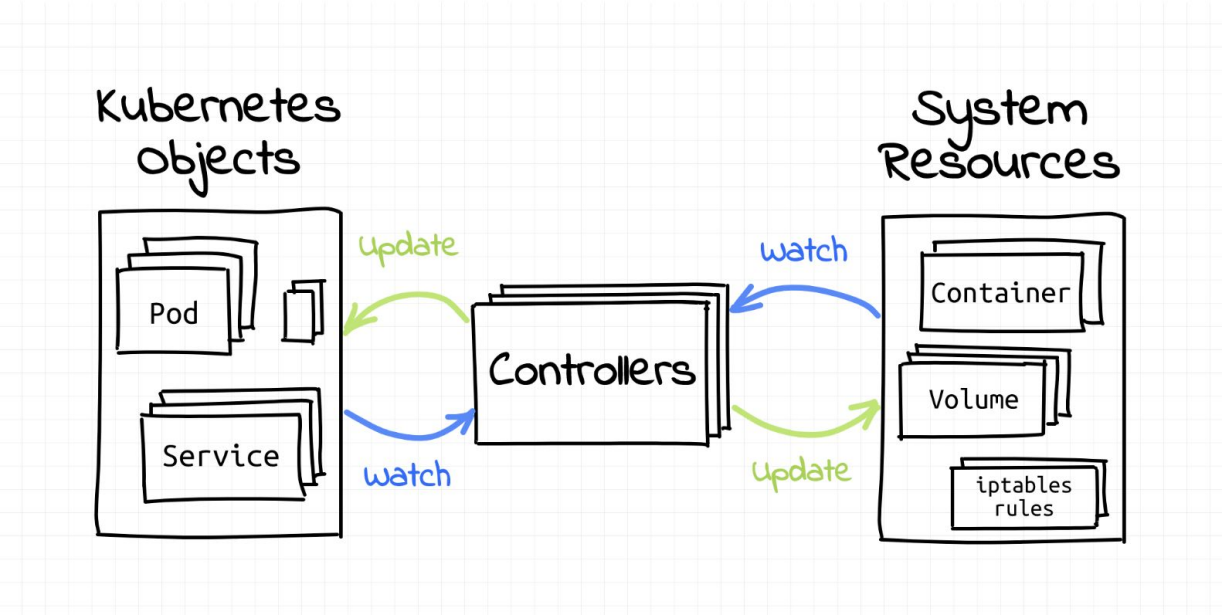
Case Study: Open AI

OpenAI adopted Kubernetes since 2016 for portability, cost saving, and improved efficiency[1,2].

Years	Nodes	Estimated Cost [3]
2018	2,500	= 3 * 2500 * 24 = US$ 180,000 / day
2021	7,500	= 3 * 2500 * 24 = US$ 540,000 / day
2023	?	

[1] https://kubernetes.io/case-studies/openai/
https://blog.pichuang.com.tw/20230214-openai-scaling-kubernetes-to-7500-nodes.html
[2] https://blog.pichuang.com.tw/20230214-openai-scaling-kubernetes-to-7500-nodes.html#_1
[3] AWS P3.2xlarge equips 8 vCores, 61 GB Memory, 1 Core of V100-16Gb GPU, charges US$ 3 per hour.

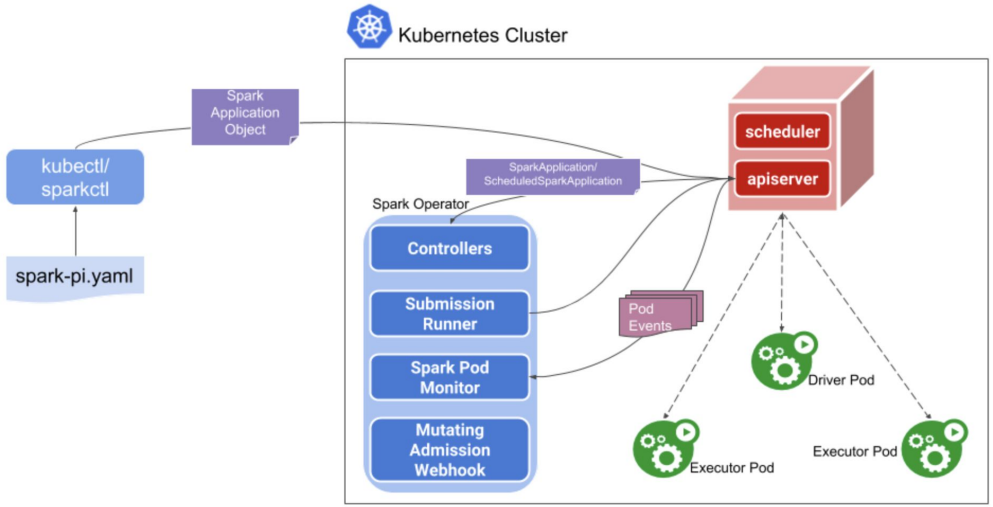
Operator pattern



https://iximiuz.com/en/posts/kubernetes-operator-pattern/

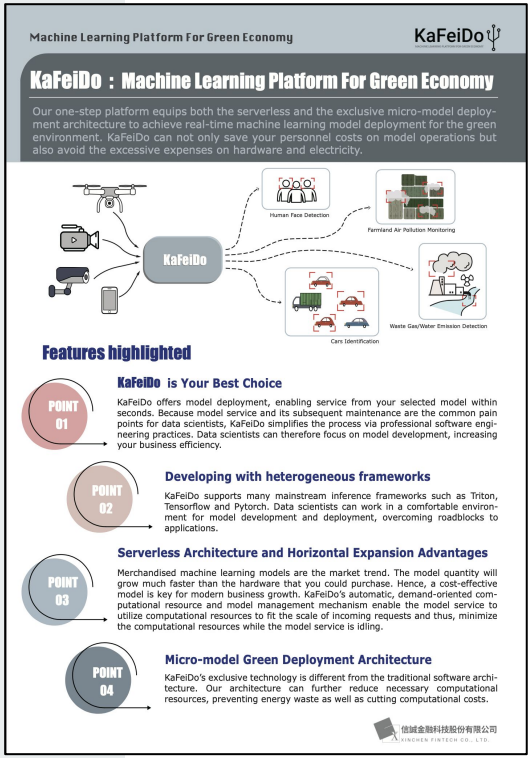
An spark CRD example

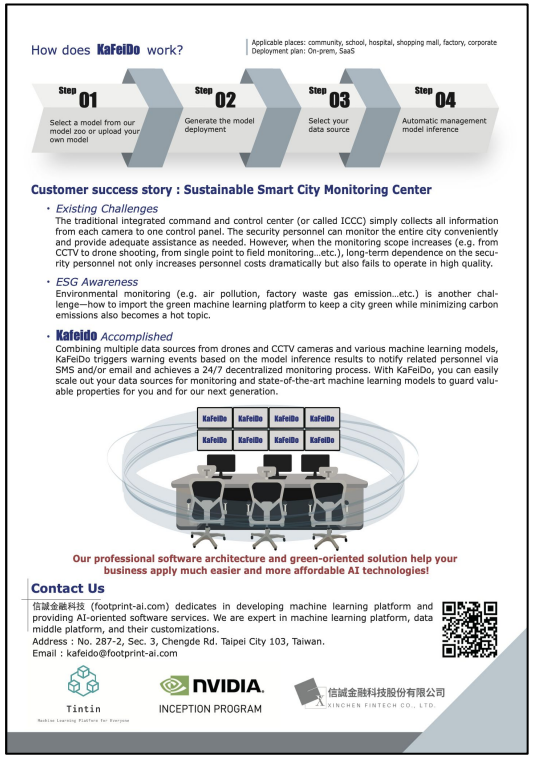
```yaml
apiVersion: "sparkoperator.k8s.io/v1beta2"
kind: SparkApplication
metadata:
  name: spark-pi
  namespace: default
spec:
  type: Scala
  mode: cluster
  image: "gcr.io/spark-operator/spark:v3.1.1"
  imagePullPolicy: Always
  mainClass: org.apache.spark.examples.SparkPi
  mainApplicationFile: "local:///opt/spark/examples/jars/spark-examples_2.12-3.1.1.jar"
  sparkVersion: "3.1.1"
  restartPolicy:
    type: Never
  volumes:
    - name: "test-volume"
      hostPath:
        path: "/tmp"
        type: Directory
```



https://github.com/GoogleCloudPlatform/spark-on-k8s-operator/
https://dzlab.github.io/ml/2020/07/14/spark-kubernetes/

Kafeido: machine learning platform for green economy

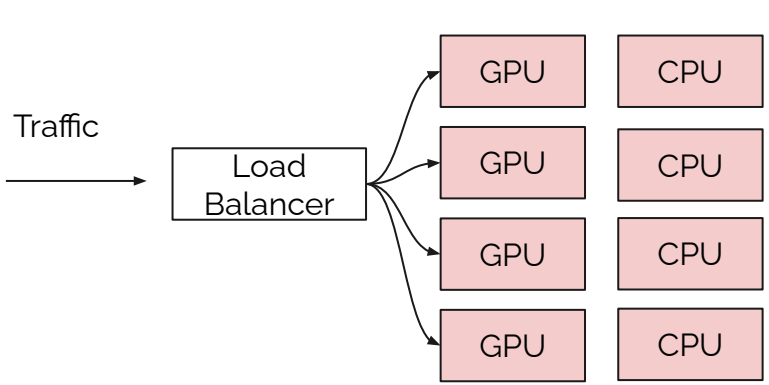


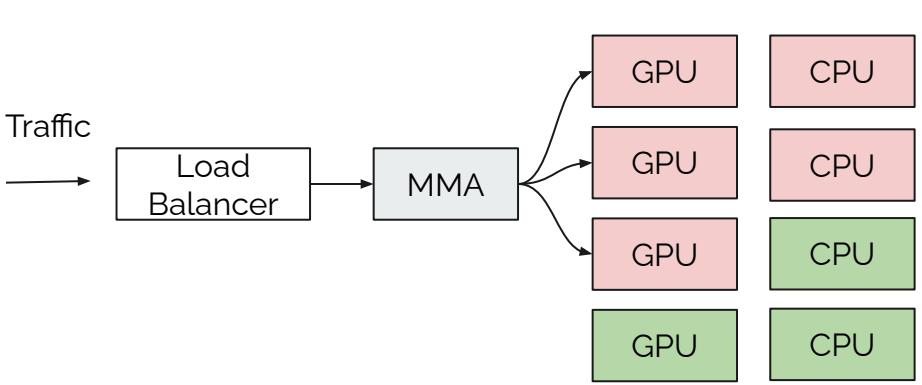


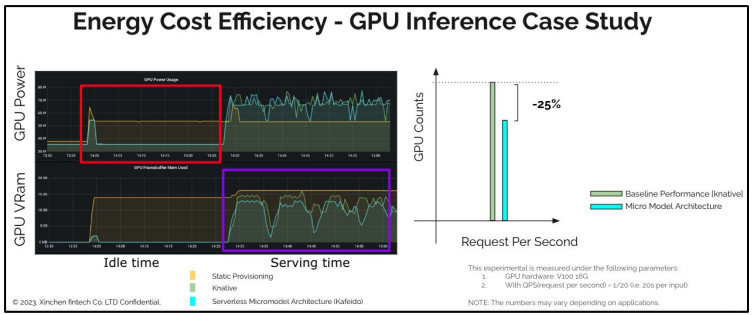
Increasing utilization with micro model architecture

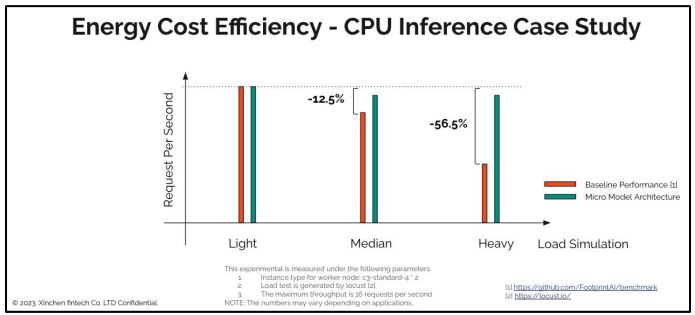
Used

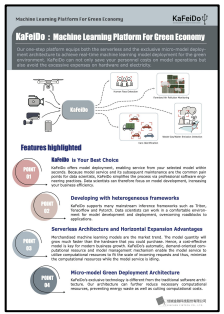
Free



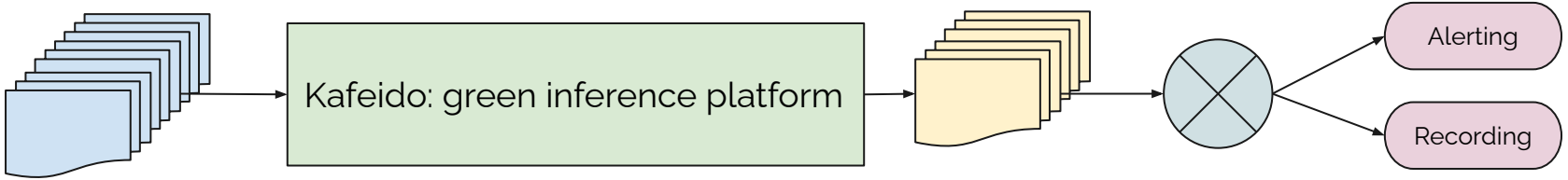








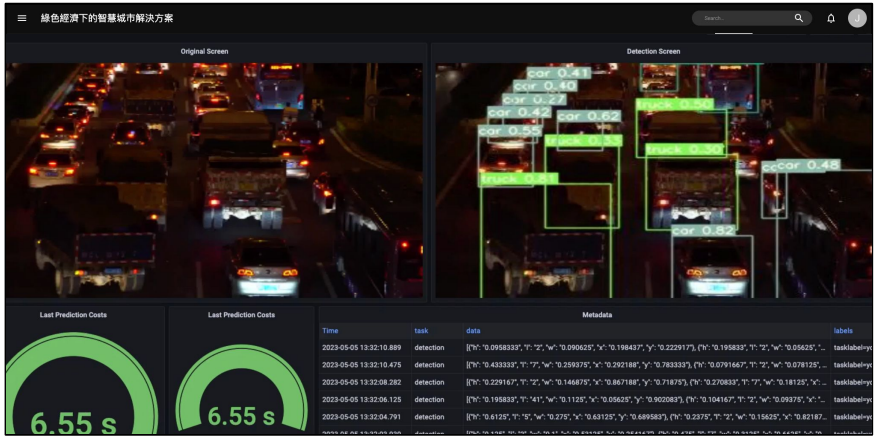
Application I - Smart Surveillance Solution for Green City [2022Q4, launched]

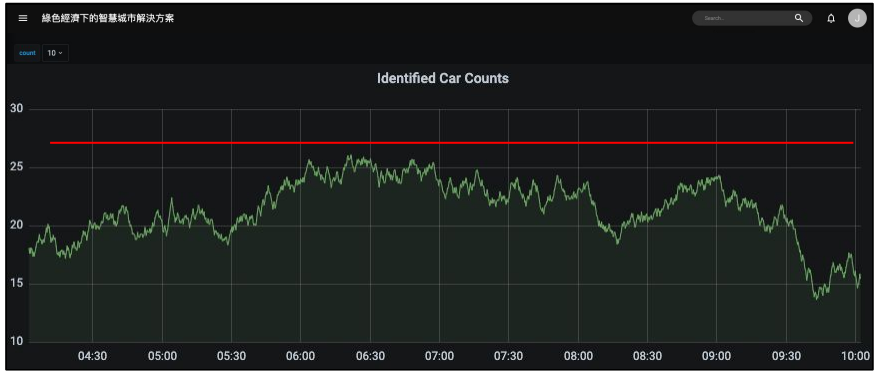


輸入資料串流
(video/text/audio)

輸出元資料串流
Metadata streaming

規則引擎
Role Engine





Parametric Insurance

For protect uncertainty when we move to sustainable solution.

https://www.bing.com/images/create/parametric-insurance-for-protect-uncertainty-when-/654ff2b8b6f148aaaae103722a52b9b7?id=MwAEAGt%2bwmf4Kn5vgEpaPg%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay

Parametric Insurance

- One variable insurance
  - Agriculture insurance where the farmer wants to protect his/her yields from rainfall over a time period
- Payout is implemented when a certain event happened in a certain time period
- I.e. *f(time period) > threshold, where f could be*
  - Weather condition
  - Bitcoin pricing
  - Cloud service status

Bitcoin price

Existing Parametric Insurance

- Weather-based index insurance
  - For protecting farmers loss due to extreme climate change.
- Renewable Energy Insurance
  - For mitigating construction and operation risk, such as Offshore wind power and photovoltaic.
- Energy Efficiency Insurance
  - For easing risks from business interruption, material damage, and so on.

We are hiring

- We use cloud-native and green technology to reduce computational cost and decarbonization.
- We use parametric insurance to protect uncertainty.

- Feel free to reach out if you have idea

https://www.bing.com/images/create/create-22we-want-you22-pictures-by-minic-the-old-sty/654ff19a6c984d5192646f11a50cdc2b?id=vYRditgOsNkH6GtTwudwFg%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay

Key takeaway

- Technology may advance, but the mindset remains constant.
- Observe the trends and pinpoint your next decade-long project.

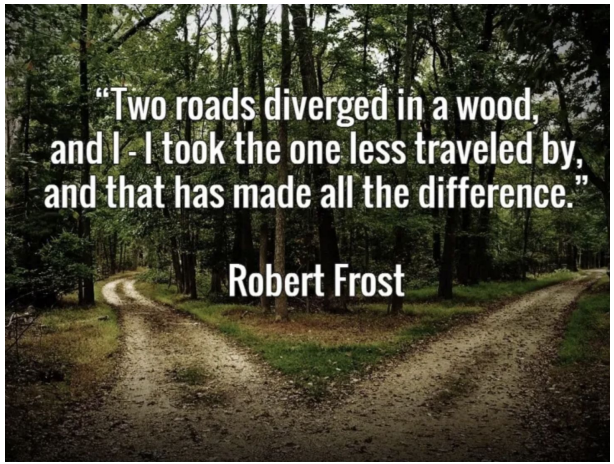


Photo ref: https://ruthpolk.com/2020/01/12/in-my-own-voice-19-two-roads-diverged-in-a-wood-and-i-i-took-the-one-less-traveled-by/

Thank You

and follow us on Facebook

Contact Us: partnership@footprint-ai.com

careers@footprint-ai.com