# GOTC

# 全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE

# OPEN SOURCE , OPEN WORLD #

## 「AI、大数据与数字经济开源技术论坛」专场

## KubeFATE: 云原生的联邦学习部署与运维平台


VMware - CTO办公室 – 云原生实验室

资深研究员

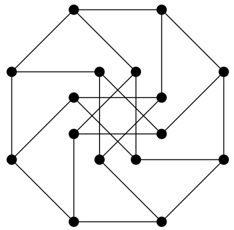
彭麟 （Layne Peng）


2021年7月10日

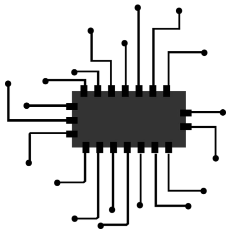
# 提纲

1. 什么联邦学习？联邦学习解决什么问题？
2. FATE: 工业级联邦学习**开源**平台；
3. **开源**云原生联邦学习方案：
    a) KubeFATE: 基于Kubernetes的联邦学习部署与运维平台
    b) FATE-Operator: Kubeflow子项目，基于KubeFATE

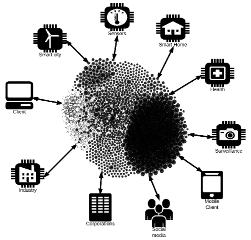
# 人工智能三大要素



算法

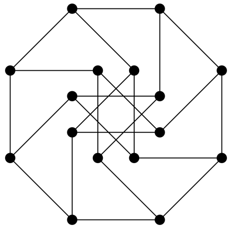


算力



数据

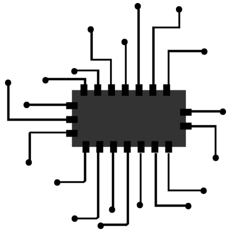
# 数据的现状并不理想


算法

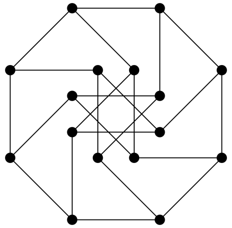

数据

数据孤岛

数据分布不均


算力

# 数据的现状并不理想

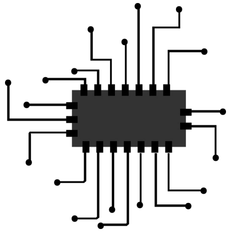


算法

数据

数据孤岛

数据分布不均



算力

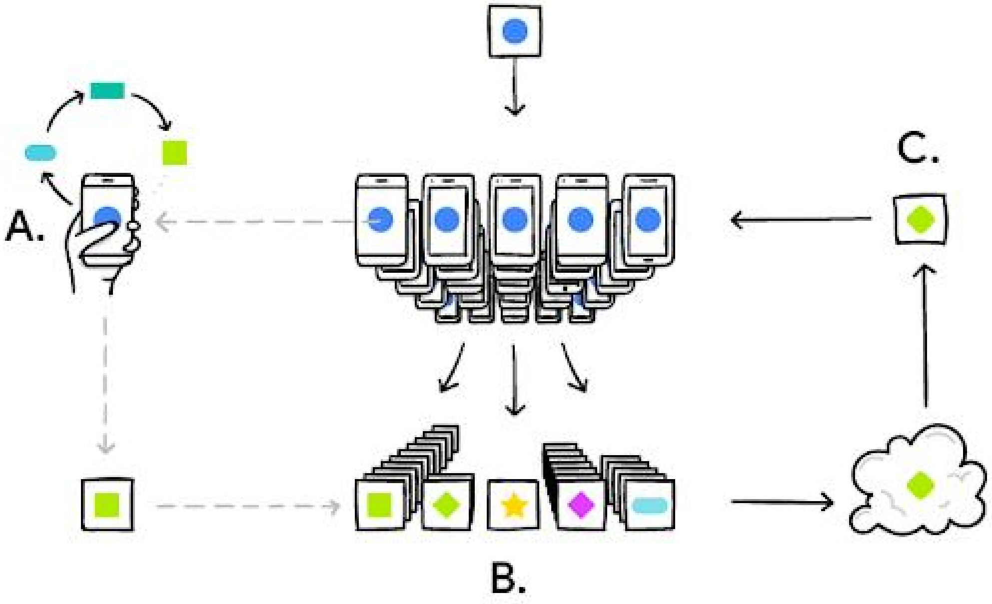
- 制造数据：*GAN*
- 利用公有(*public*)和开放(*open*)数据：迁移学习
- 私有数据方合作一起训练: 联邦学习 (*Federated learning*)

# 联邦学习概念出现

数据

数据孤岛

数据分布不均




(Source: Federated Learning: Collaborative Machine Learning without Centralized Training Data, Google AI Blog, 2017）

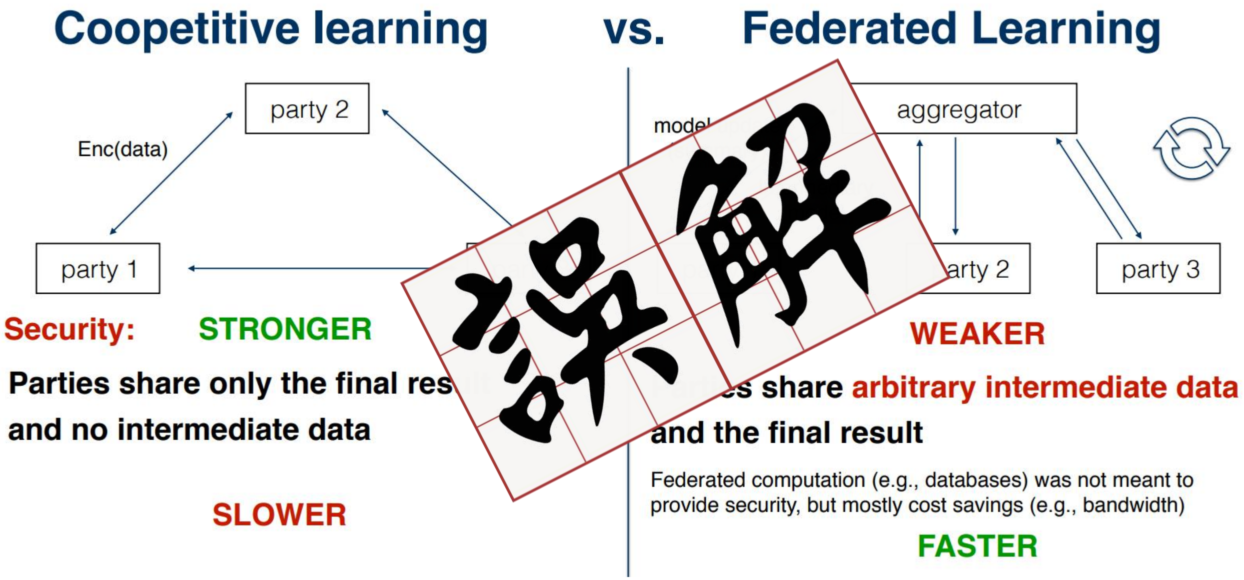
# 联邦学习的误解：无隐私保护

早期的研究报告、论文往往基于无隐私保护的联邦学习方案。

数据孤岛

数据分布不均

数据




(Source: Secure Collaborative Learning, 2017)

# （安全&保护隐私的）联邦学习

数据

数据孤岛

数据分布不均

隐私法律法规

数据安全

。。。


- 联邦学习 (Federated learning) => （**安全&保护隐私的**）联邦学习

# （安全&保护隐私的）联邦学习

数据

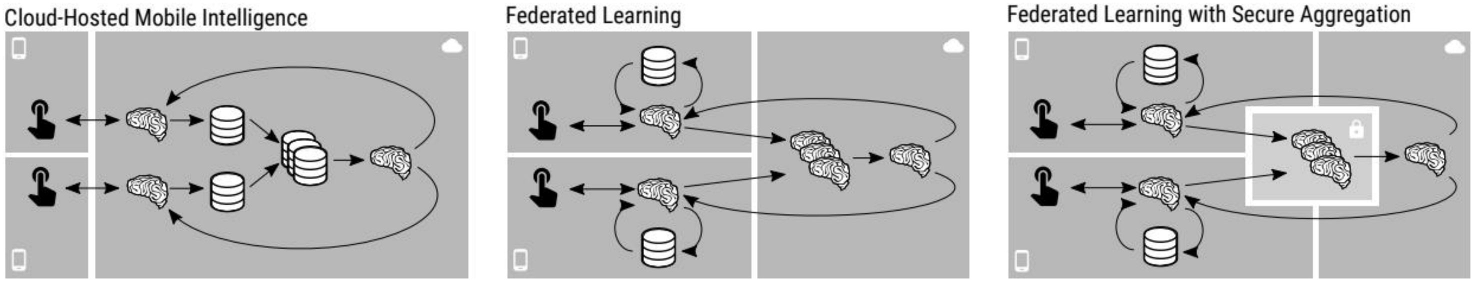
数据孤岛

数据分布不均

隐私法律法规

数据安全

。 。 。



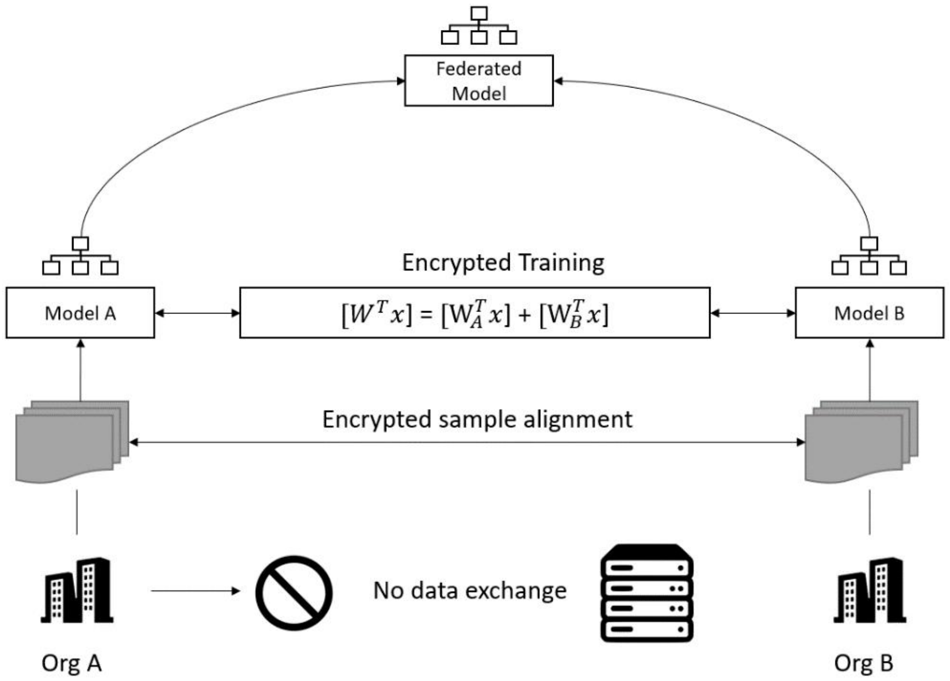

(Source: Practical Secure Aggregation
for Privacy-Preserving Machine Learning, Keith Bonawitz et al, 2017)


# 联邦学习的定义

数据

- 数据孤岛
- 数据分布不均
- 隐私法律法规
- 数据安全

。。。


- 两个或更多的（子）组织共同训练模型
- 组织间无数据交换
- **加密模型在多方安全计算框架下共同训练：**
  - 同态加密
  - 共享密钥
  - 不经意传输
  - …

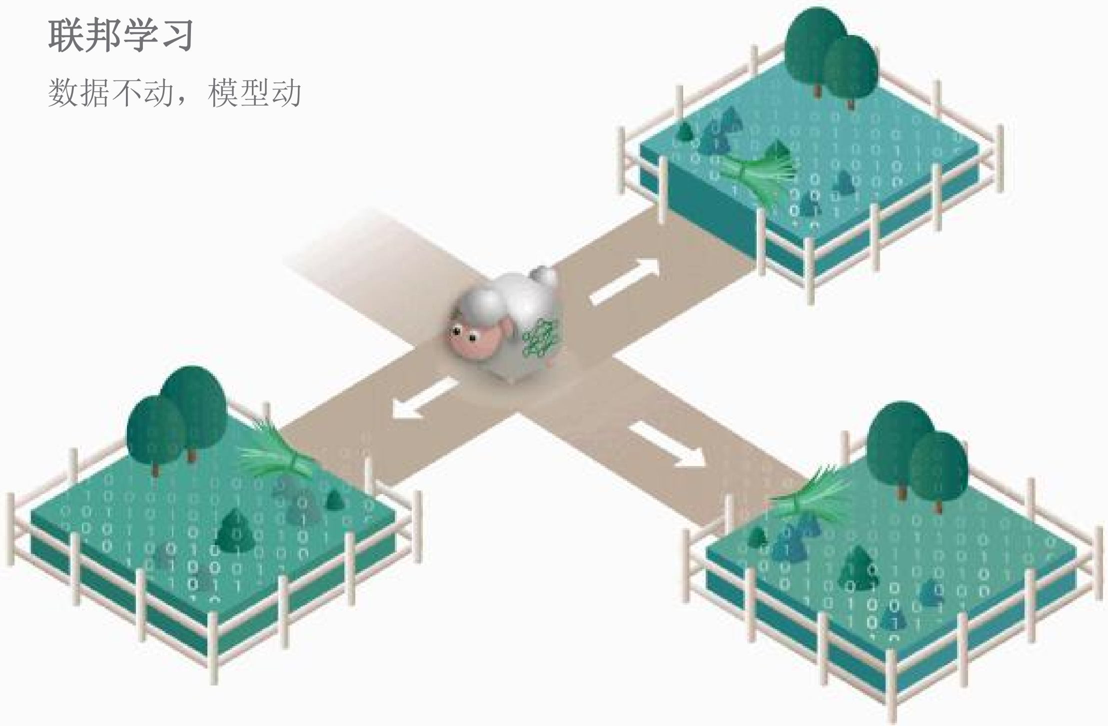
# 联邦学习与传统的机器学习

**传统的机器学习**

模型不动，数据动

**联邦学习**

数据不动，模型动



(Source: *Federated Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning) , Qiang yang , et al.* )

## 数据不动模型动，数据可用不可见

vmware®

# 联邦学习是解决数据孤岛问题的一个可行方案

## Advances and Open Problems in Federated Learning


Peter Kairouz[7*] H. Brendan McMahan[7*] Brendan Avent[21] Aurélien Bellet[9]
Mehdi Bennis[19] Arjun Nitin Bhagoji[13] Keith Bonawitz[7] Zachary Charles[7]
Graham Cormode[23] Rachel Cummings[6] Rafael G.L. D'Oliveira[14]
Salim El Rouayheb[14] David Evans[22] Josh Gardner[24] Zachary Garrett[7]
Adrià Gascón[7] Badih Ghazi[7] Phillip B. Gibbons[2] Marco Gruteser[7,14]
Zaid Harchaoui[24] Chaoyang He[21] Lie He [4] Zhouyuan Huo [20]
Ben Hutchinson[7] Justin Hsu[25] Martin Jaggi[4] Tara Javidi[17] Gauri Joshi[2]
Mikhail Khodak[2] Jakub Konečný[7] Aleksandra Korolova[21] Farinaz Koushanfar[17]
Sanmi Koyejo[7,18] Tancrède Lepoint[7] Yang Liu[12] Prateek Mittal[13]
Mehryar Mohri[7] Richard Nock[1] Ayfer Özgür[15] Rasmus Pagh[7,10]
Mariana Raykova[7] Hang Qi[7] Daniel Ramage[7] Ramesh Raskar[11]
Dawn Song[16] Weikang Song[7] Sebastian U. Stich[4] Ziteng Sun[3]
Ananda Theertha Suresh[7] Florian Tramèr[15] Praneeth Vepakomma[11] Jianyu Wang[2]
Li Xiong[5] Zheng Xu[7] Qiang Yang[8] Felix X. Yu[7] Han Yu[12] Sen Zhao[7]

[1]Australian National University, [2]Carnegie Mellon University, [3]Cornell University,
[4]École Polytechnique Fédérale de Lausanne, [5]Emory University, [6]Georgia Institute of Technology,
[7]Google Research, [8]Hong Kong University of Science and Technology, [9]INRIA, [10]IT University of Copenhagen,
[11]Massachusetts Institute of Technology, [12]Nanyang Technological University, [13]Princeton University,
[14]Rutgers University, [15]Stanford University, [16]University of California Berkeley,
[17] University of California San Diego, [18]University of Illinois Urbana-Champaign, [19]University of Oulu,
[20]University of Pittsburgh, [21]University of Southern California, [22]University of Virginia,
[23]University of Warwick, [24]University of Washington, [25]University of Wisconsin–Madison





**Abstract**

Federated learning (FL) is a machine learning setting where many clients (e.g. mobile devices or

# 联邦学习是解决数据孤岛问题的一个可行方案

arXiv:1912.04977v1 [cs.LG] 10 Dec 2019

## Advances and Open Problems in Federated Learning

Peter Kairouz[7*] H. Brendan McMahan[7*] Brendan Avent[21] Aurélien Bellet[9] Mehdi Bennis[19] Arjun Nitin Bhagoji[13] Keith Bonawitz[7] Zachary Charles[7] Graham Cormode[23] Rachel Cummings[6] Rafael G.L. D'Oliveira[14] Salim El Rouayheb[14] David Evans[22] Josh Gardner[24] Zachary Garrett[7] Adrià Gascón[7] Badih Ghazi[7] Phillip B. Gibbons[2] Marco Gruteser[7,14] Zaid Harchaoui[24] Chaoyang He[21] Lie He [4] Zhouyuan Huo [20] Ben Hutchinson[7] Justin Hsu[25] Martin Jaggi[4] Tara Javidi[17] Gauri Joshi[2] Mikhail Khodak[2] Jakub Konečný[7] Aleksandra Korolova[21] Farinaz Koushanfar[17] Sanmi Koyejo[7,18] Tancrède Lepoint[7] Yang Liu[12] Prateek Mittal[13] Mehryar Mohri[7] Richard Nock[1] Ayfer Özgür[15] Rasmus Pagh[7,10] Mariana Raykova[7] Hang Qi[7] Daniel Ramage[7] Ramesh Raskar[11] Dawn Song[16] Weikang Song[7] Sebastian U. Stich[4] Ziteng Sun[3] Ananda Theertha Suresh[7] Florian Tramèr[15] Praneeth Vepakomma[11] Jianyu Wang[2] Li Xiong[5] Zheng Xu[7] Qiang Yang[8] Felix X. Yu[7] Han Yu[12] Sen Zhao[7]

[1]Australian National University, [2]Carnegie Mellon University, [3]Cornell University, [4]École Polytechnique Fédérale de Lausanne, [5]Emory University, [6]Georgia Institute of Technology, [7]Google Research, [8]Hong Kong University of Science and Technology, [9]INRIA, [10]IT University of Copenhagen, [11]Massachusetts Institute of Technology, [12]Nanyang Technological University, [13]Princeton University, [14]Rutgers University, [15]Stanford University, [16]University of California Berkeley, [17] University of California San Diego, [18]University of Illinois Urbana-Champaign, [19]University of Oulu, [20]University of Pittsburgh, [21]University of Southern California, [22]University of Virginia, [23]University of Warwick, [24]University of Washington, [25]University of Wisconsin–Madison

**Abstract**

Federated learning (FL) is a machine learning setting where many clients (e.g. mobile devices or

ICS 35.240.40
CCS A 11

**JR**

中 华 人 民 共 和 国 金 融 行 业 标 准

JR/T 0196—2020

多方安全计算金融应用技术规范

Secure multi-party computation financial application technical specification

2020－11－24 发布 2020－11－24 实施

中国人民银行 发 布

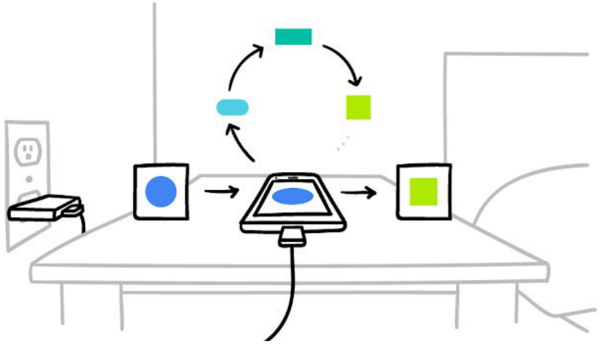

CAICT
中国信息通信研究院
China Academy of Information and Communications Technology

北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

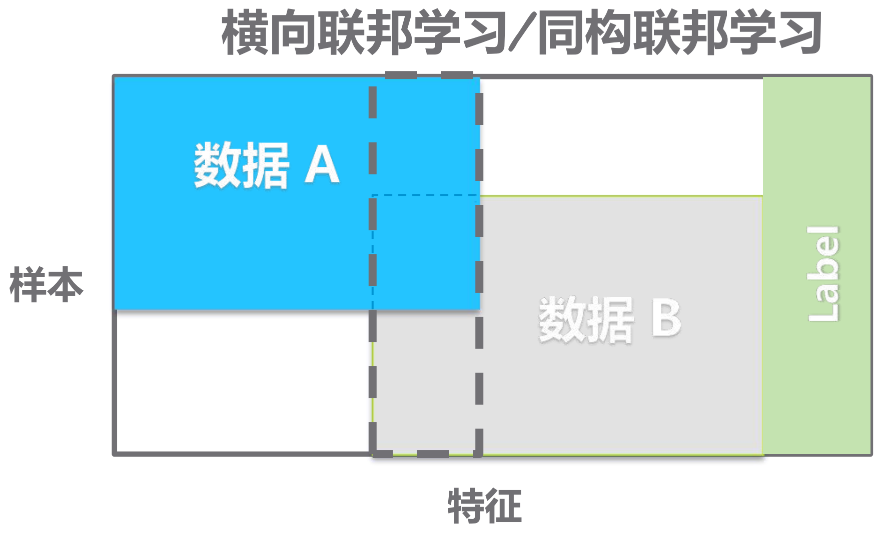
# 联邦学习的分类

数据孤岛情况 1：样例分散在不同的组织，单个组织样例不足以支持优质训练。。。

样本分区的联邦学习

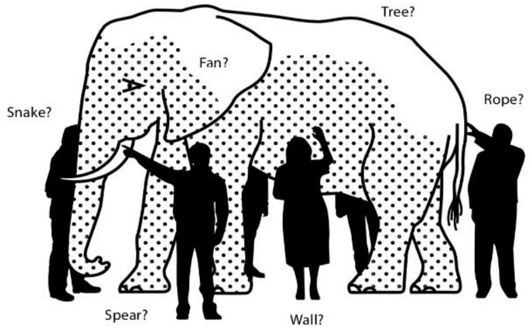


(Source: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html)

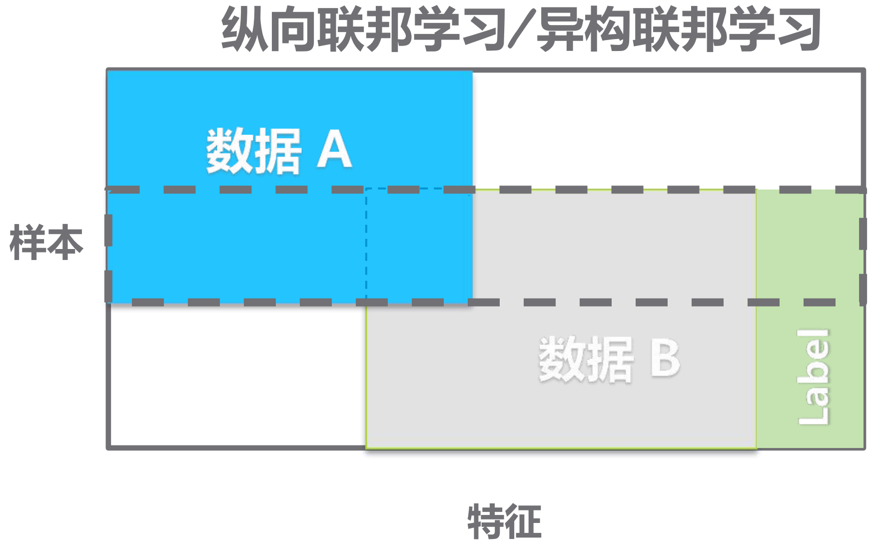



数据孤岛情况 2：样本数据的特征分散在不同组织，单个组织有样本片面的理解，造成训练结果偏差。。。

特征分区的联邦学习




(Source：中国寓言，盲人摸象)

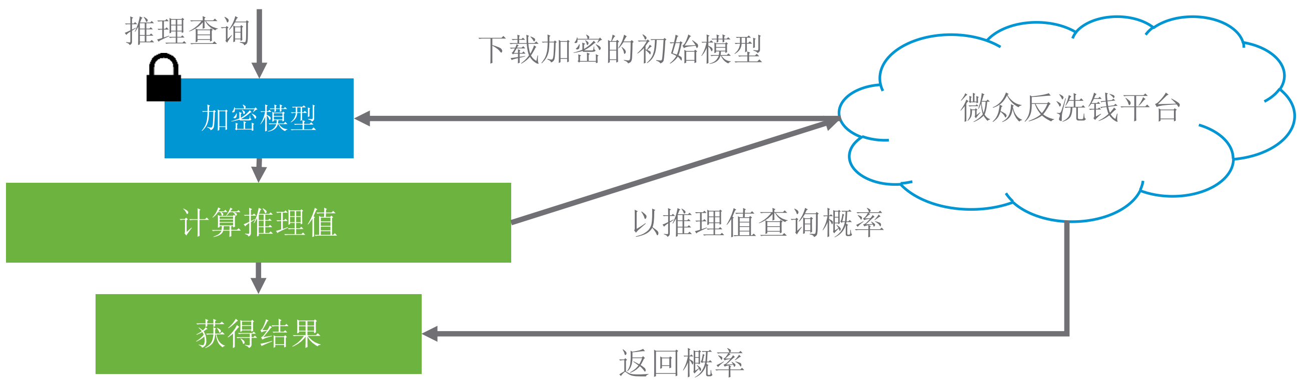


# 横向、纵向联邦学习的案例

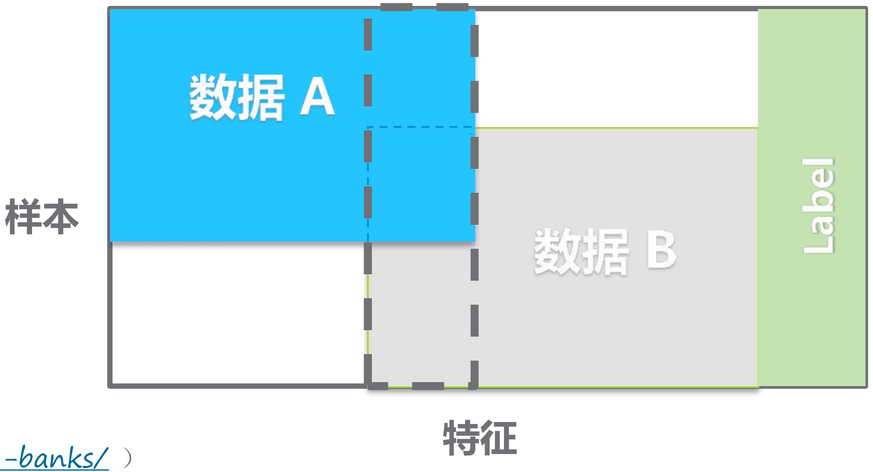
## 跨银行反洗钱应用

在符合法规情况下，组合各银行数据成大量实证案例

AUC增加了14%




横向联邦学习/同构联邦学习

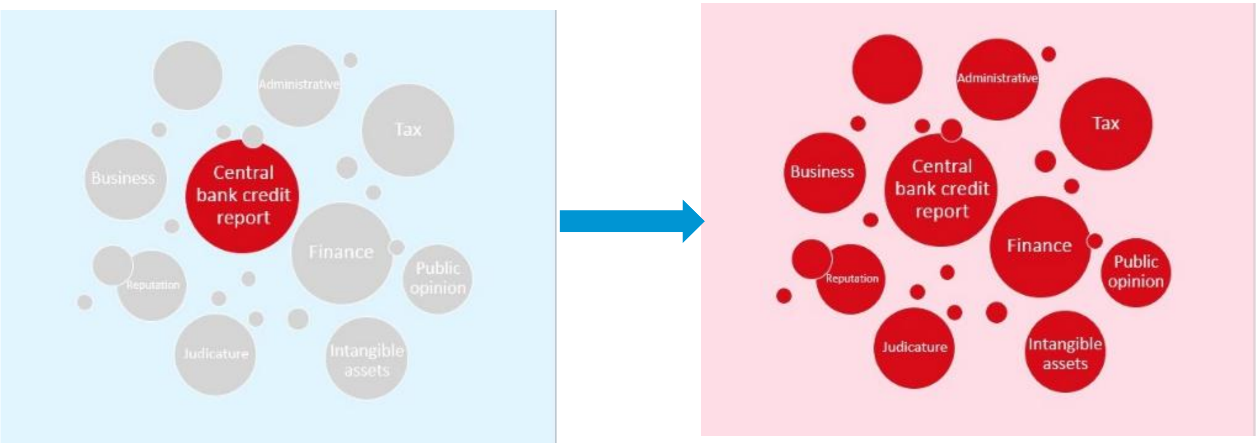



(Source: https://www.fedai.org/cases/utilization-of-fate-in-anti-money-laundering-through-multiple-banks/ )

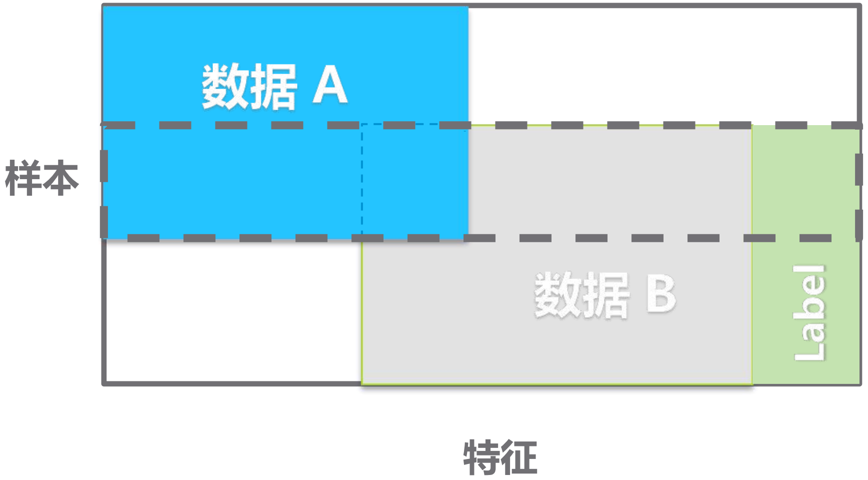
## 小微企业信用风险管理

多元数据来源组合获得更准确的用户画像

AUC增加了12%




纵向联邦学习/异构联邦学习




(Source: https://www.fedai.org/cases/utilization-of-fate-in-risk-management-of-credit-in-small-and-micro-enterprises/)

# FATE: Federated AI Technology Enabler

# FATE: Federated AI Technology Enabler


FATE是开箱即用的联邦学习平台：

1. 内置典型的联邦学算法；
2. 可视化建模界面；
3. DAG工作流引擎；
4. 支持多种多方计算安全协议：同态加密、共享密钥，*etc.*
5. 支持审计等功能，满足银监等保要求；
6. 分布式计算、存储、传输引擎；
7. 支持异构加速器。




1. 开箱即用的算法；
2. 联邦学习算法开发框架：
   a) 底层工具
   b) 通信协议引擎
   c) 工作流引擎
   *d) 互联互通协议*
   *e) 算法编译器*

联邦算法

框架

驱动环境

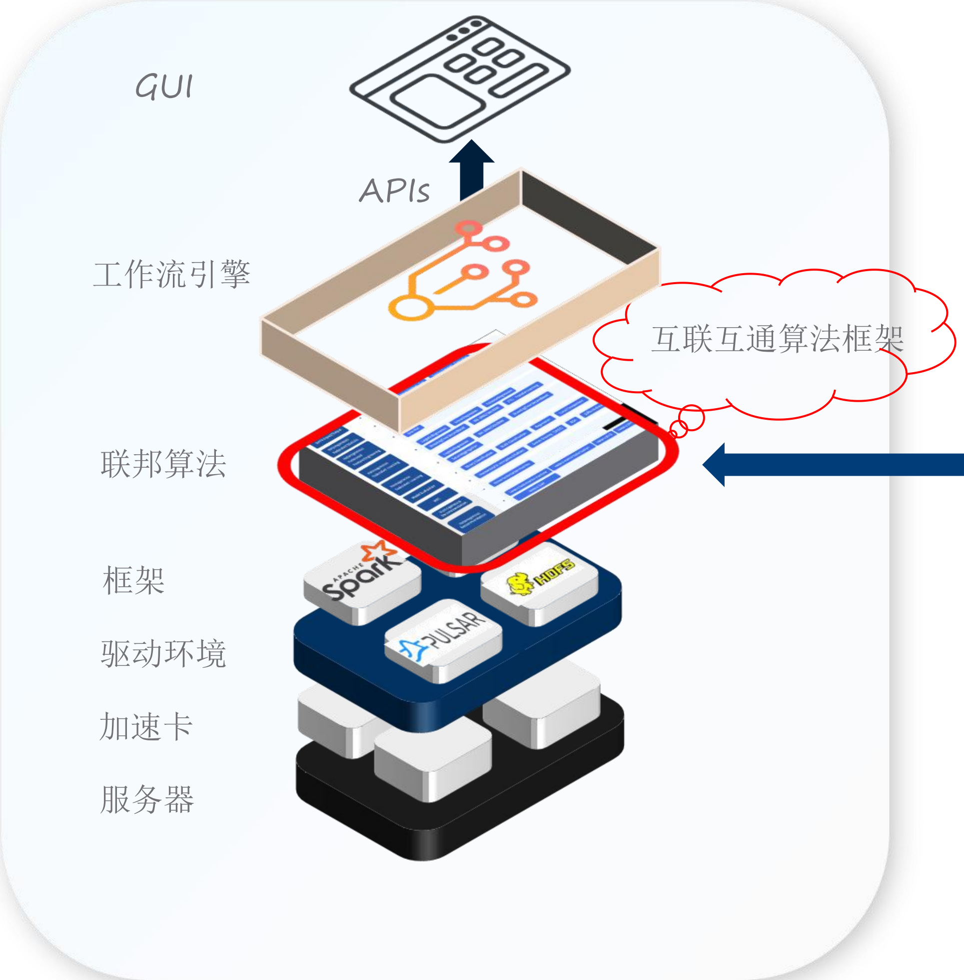
加速卡

服务器


1. 重用已有算力：支持开源计算、传输、存储框架
   a) Spark
   b) Pulsar/RabbitMQ
   c) HDFS
   d) Hive
   e) …
2. 异构加速器：
   a) GPU
   b) FPGA
   c) ARM

# FATE: Federated AI Technology Enabler v1.7.0

FATE v1.7.0是一个联邦学习的生态系统（FedAI）



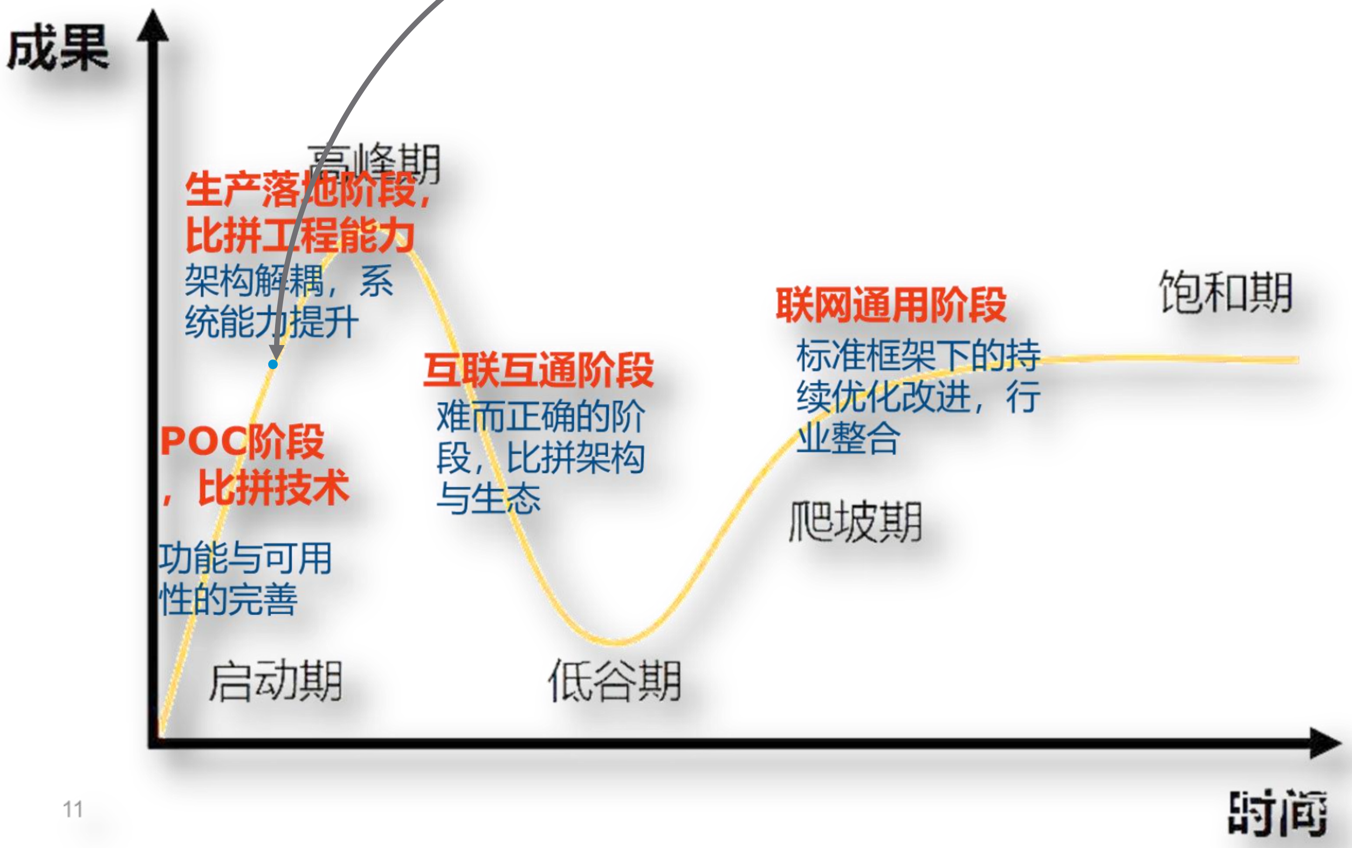

算法市场

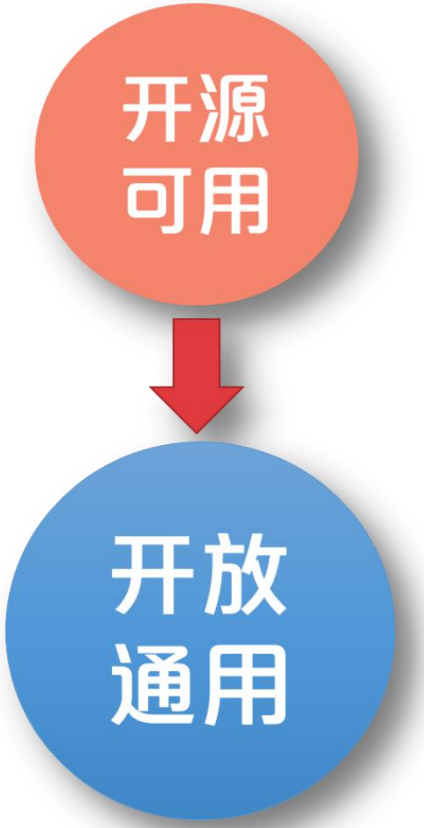

Source: [破解不同技术平台交互阻碍，「富数科技」和「微众银行」实现异构联邦学习平台互通]

# 联邦学习的发展

我们在这里：

1. 为下一阶段搭建互联互通的平台
2. 推动更多的生产落地。



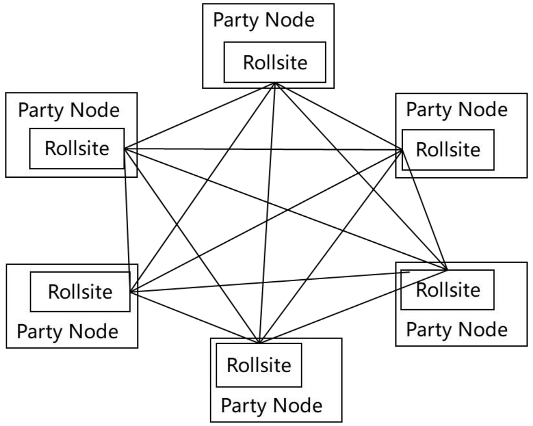




Source:企业级联邦学习平台建设的探索与思考，中国银联金融科技研究院，周雍恺

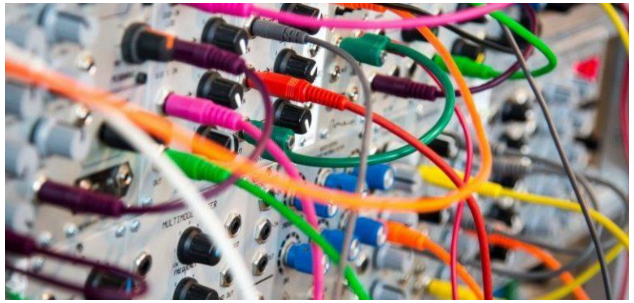
# FATE设计为工业级联邦学习开源平台，但是。。。

## 架构及部署环境复杂

1. 分布式系统、分层结构

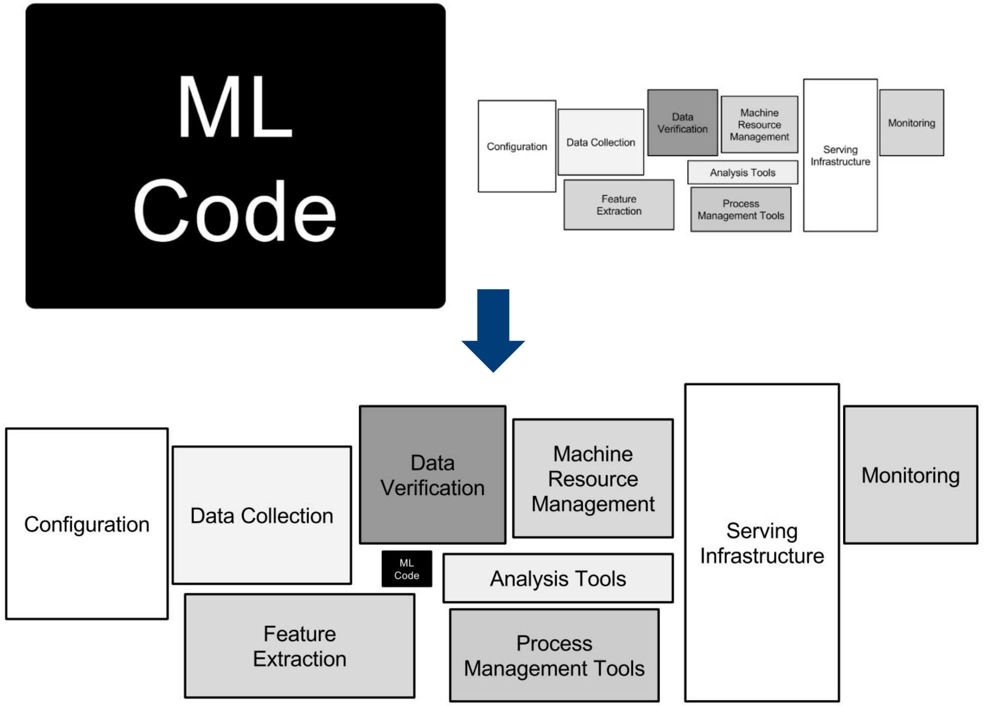



2. 复杂的企业环境：安全、网络、遗留系统适配



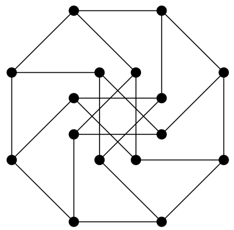
## 机器学习是一个系统工程

1. 联邦学习需要与已有系统对接
2. 联邦学习需要管理功能：数据、权限、*etc.*

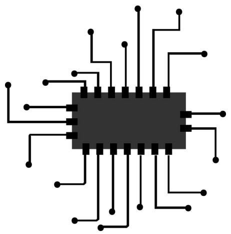



(Source: *Hidden Technical Debt in Machine Learning Systems, D. Sculley, et al.*)

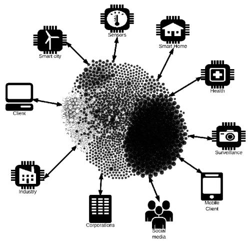
# 人工智能第四要素



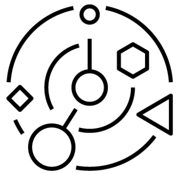
算法



算力

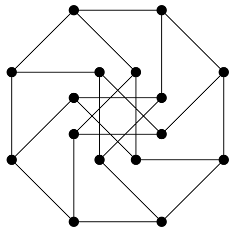


数据

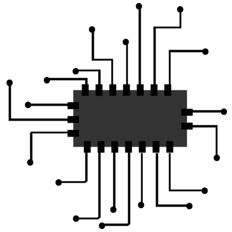


运维

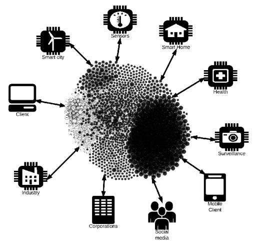
# 人工智能第四要素



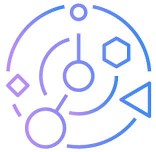
算法



算力



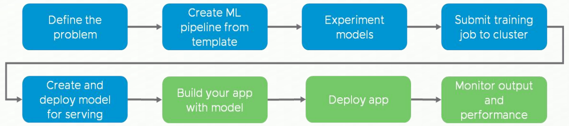
数据



云原生联邦学习

可插拔

可扩展




全生命周期管理

安全

管理

高可用

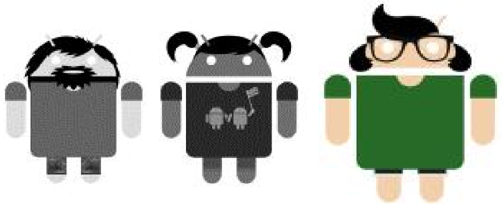
# KubeFATE: 云原生联邦学习平台

KubeFATE

Docker Compose

1. 测试、体验多方FATE集群；
2. 上手简单。

vs.



ML Expert SWE SRE

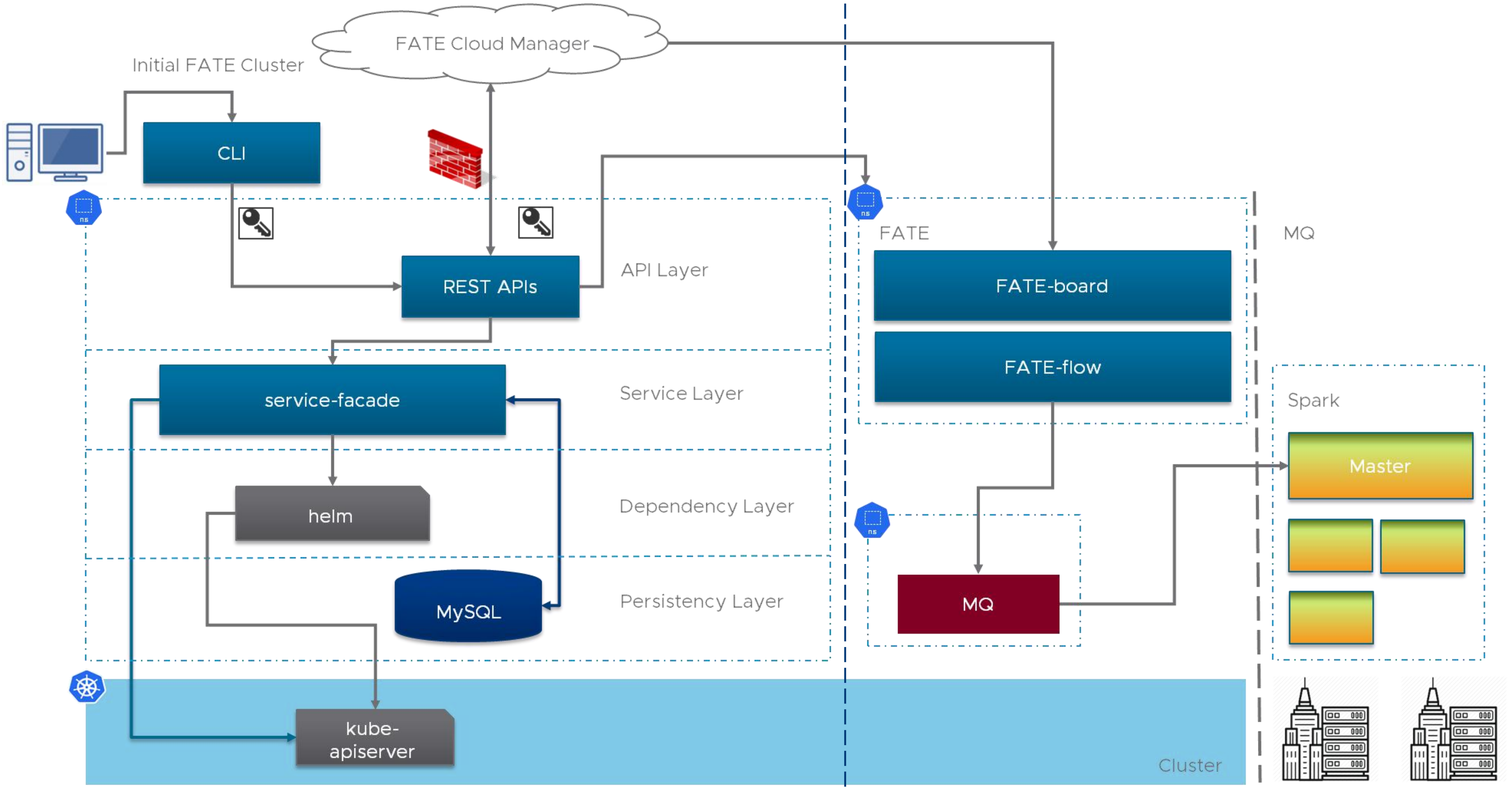
kubernetes

1. 面向生产环境：
   1) 支持多个FATE环境及集群；
   2) 声明式扩展能力；
   3) 升级，迁移；
   4) 日志及监控功能
2. 强大的定制功能

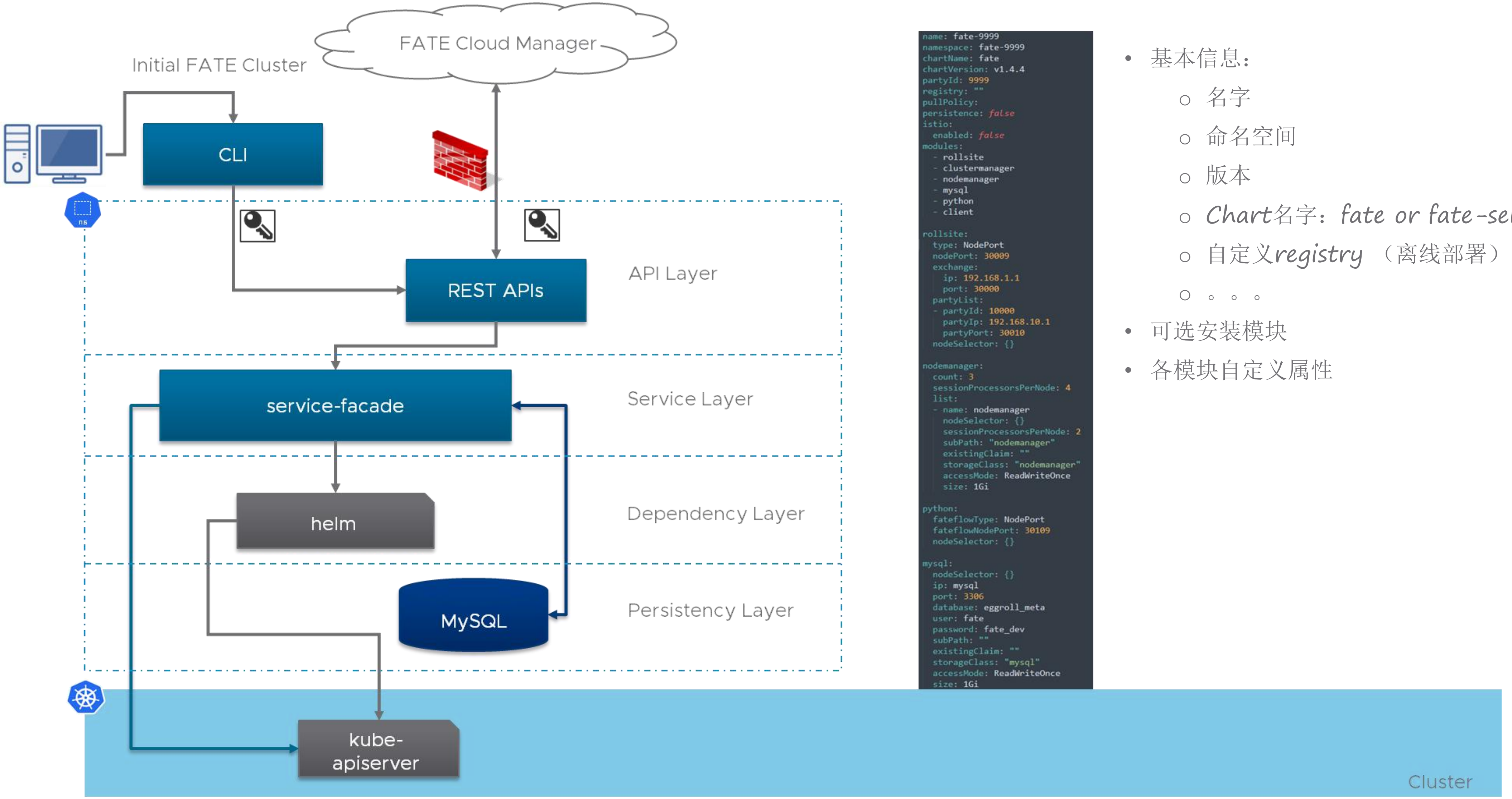
Containers

amazon web services | Microsoft Azure | IBM SOFTLAYER | Google Cloud | VMware Tanzu

# KubeFATE: 架构、模块



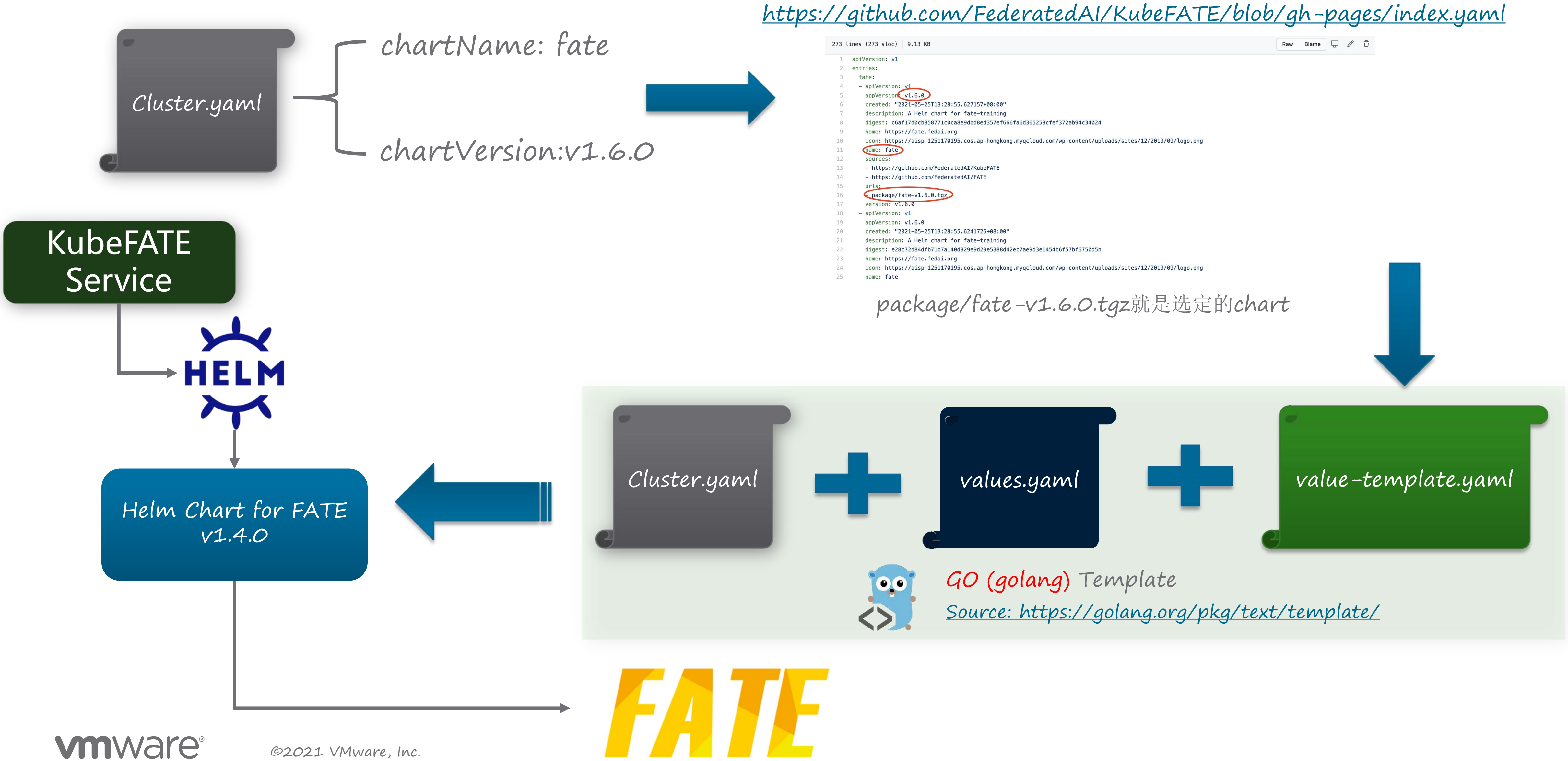
# KubeFATE: cluster.yaml



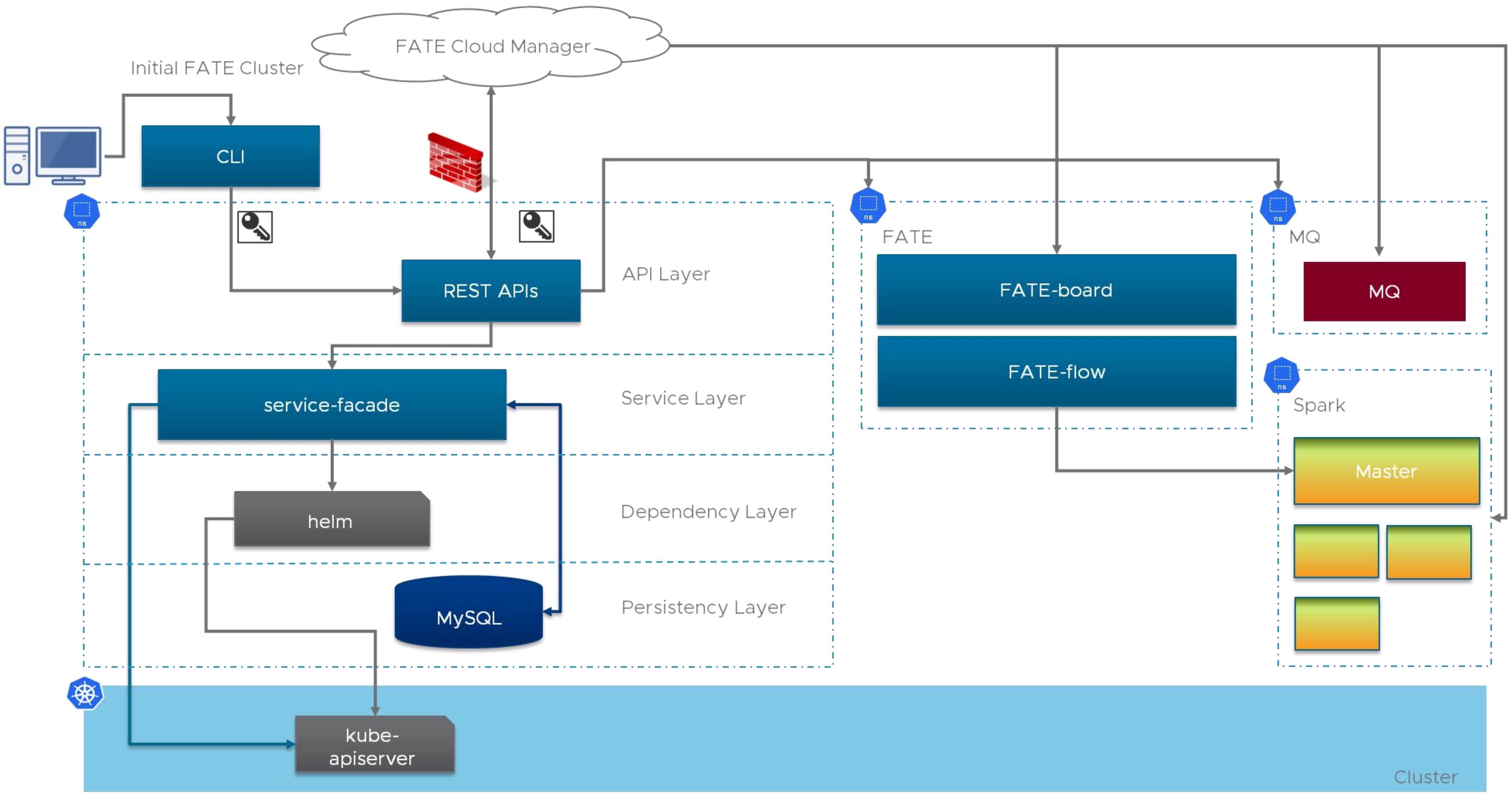

```yaml
name: fate-9999
namespace: fate-9999
chartName: fate
chartVersion: v1.4.4
partyId: 9999
registry: ""
pullPolicy:
persistence: false
istio:
  enabled: false
modules:
  - rollsite
  - clustermanager
  - nodemanager
  - mysql
  - python
  - client

rollsite:
  type: NodePort
  nodePort: 30009
  exchange:
    ip: 192.168.1.1
    port: 30000
  partyList:
  - partyId: 10000
    partyIp: 192.168.10.1
    partyPort: 30010
  nodeSelector: {}

nodemanager:
  count: 3
  sessionProcessorsPerNode: 4
  list:
  - name: nodemanager
    nodeSelector: {}
    sessionProcessorsPerNode: 2
    subPath: "nodemanager"
    existingClaim: ""
    storageClass: "nodemanager"
    accessMode: ReadWriteOnce
    size: 1Gi

python:
  fateflowType: NodePort
  fateflowNodePort: 30109
  nodeSelector: {}

mysql:
  nodeSelector: {}
  ip: mysql
  port: 3306
  database: eggroll_meta
  user: fate
  password: fate_dev
  subPath: ""
  existingClaim: ""
  storageClass: "mysql"
  accessMode: ReadWriteOnce
  size: 1Gi
```

- 基本信息：
  - 名字
  - 命名空间
  - 版本
  - Chart名字：fate or fate-serving
  - 自定义registry （离线部署）
  - 。。。
- 可选安装模块
- 各模块自定义属性

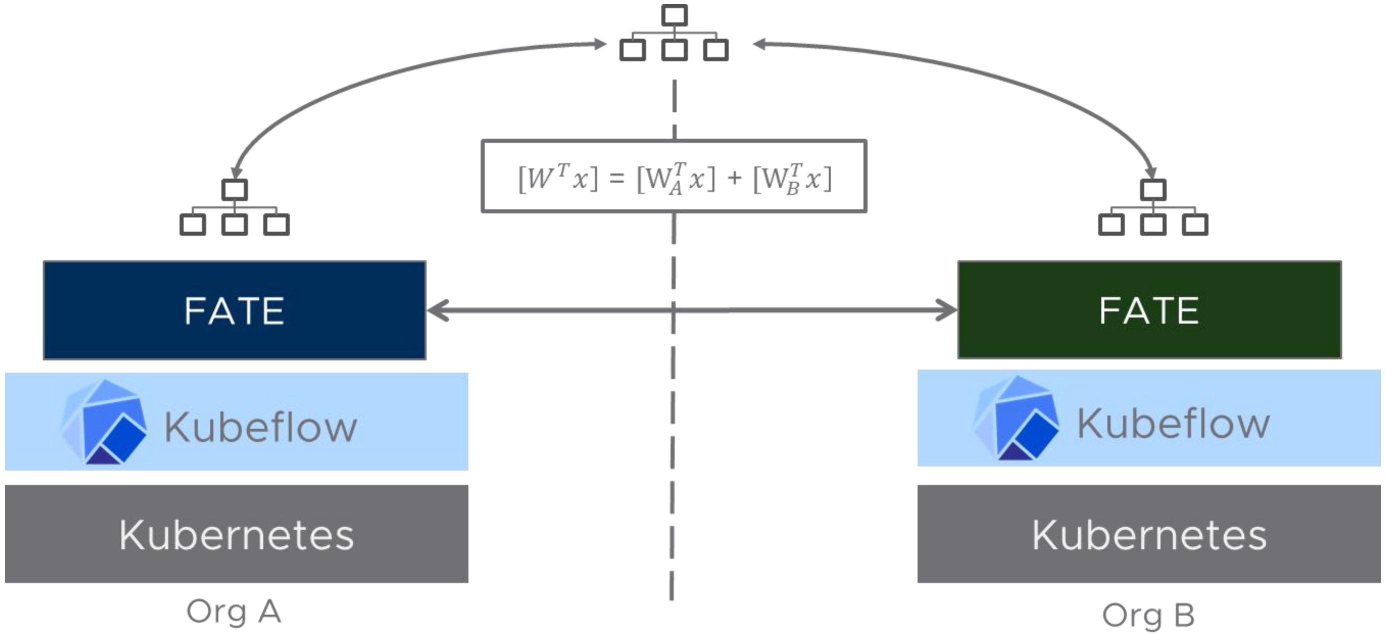
# KubeFATE: chart渲染



# KubeFATE: 定制化部署



# FATE-Operator: Kubeflow官方子项目，Kubeflow联邦学习方案




1. 如何快速、简单、按需部署分布式的FATE集群？
2. 如何使用Kubeflow的工作流引擎提交批量的联邦学习训练任务？
3. 如何与Kubeflow已有的生态更好整合，提供端对端的联邦学习生命周期支持？

FATE-Operator: https://github.com/kubeflow/fate-operator

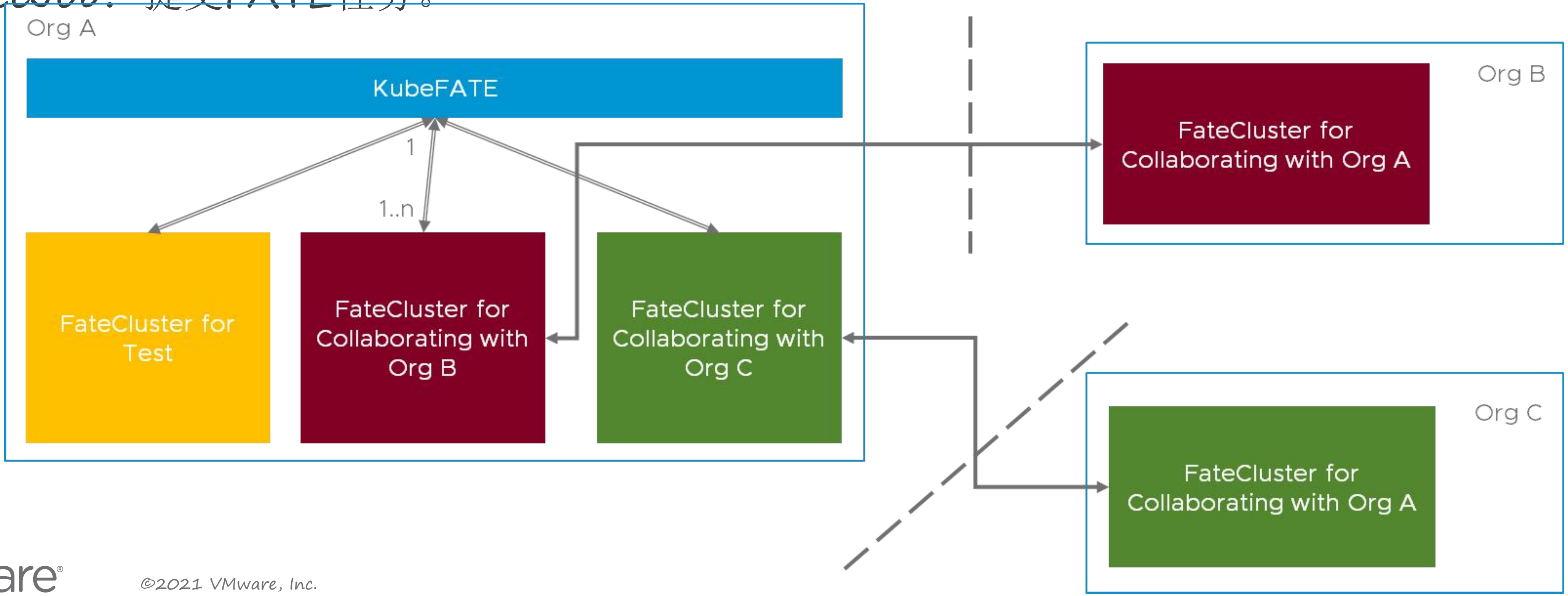
# FATE-Operator: Kubeflow官方子项目，Kubeflow联邦学习方案

基于Kubebuilder (https://github.com/kubernetes-sigs/kubebuilder). 提供了三个自定义资源：

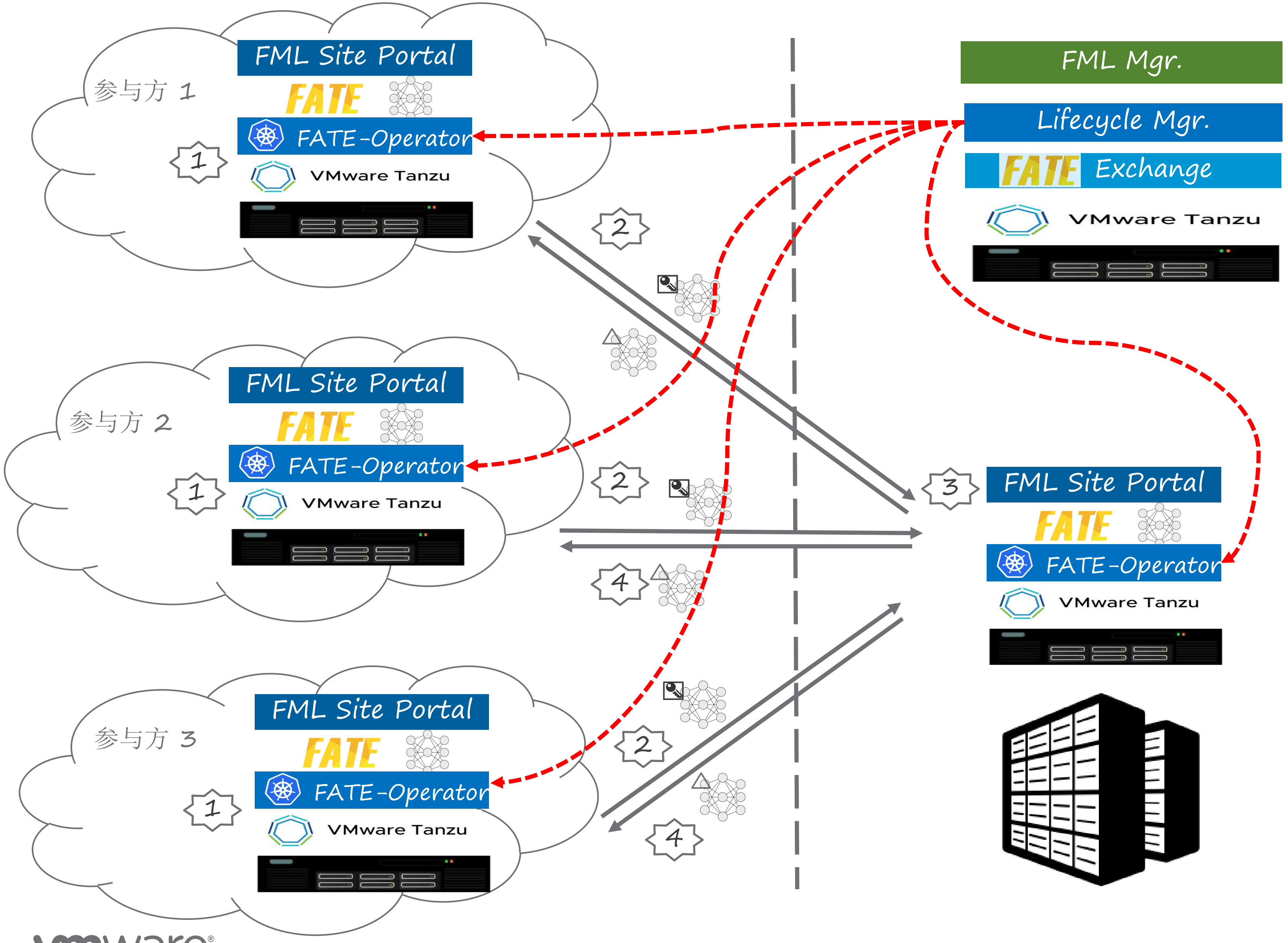
1. KubeFATE: 部署KubeFATE项目；

2. FateCluster：部署FATE集群；

3. FateJob：提交FATE任务。



# KubeFATE: FATE+VCF企业级方案




联邦训练管理：联邦数据管理、模型管理、授权。。。

生命周期管理：部署，联邦建立，监控等等

基于VCF的HA，安全方案

# 总结

1. 联邦学习是解决小数据、数据孤岛的可行方案，核心是“数据不动模型动”；
2. FATE是面向生产的开源联邦学习平台，并且在下一个版本更开放。欢迎大家的试用与贡献；
3. 联邦学习的复杂性提出了运维的需求，为此我们提出云原生联邦学习的概念，并开源：
   - KubeFATE： https://github.com/FederatedAI/KubeFATE
   - FATE-Operator： https://github.com/kubeflow/fate-operator
   - FATE: https://github.com/FederatedAI/FATE

VMware中国研发中心

回复“kubefate”加入KubeFATE交流群

FATE联邦学习技术交流群

# THANKS