innovation
hub sessions

Organizado por:

pue

pue

9:30 a 10:15h

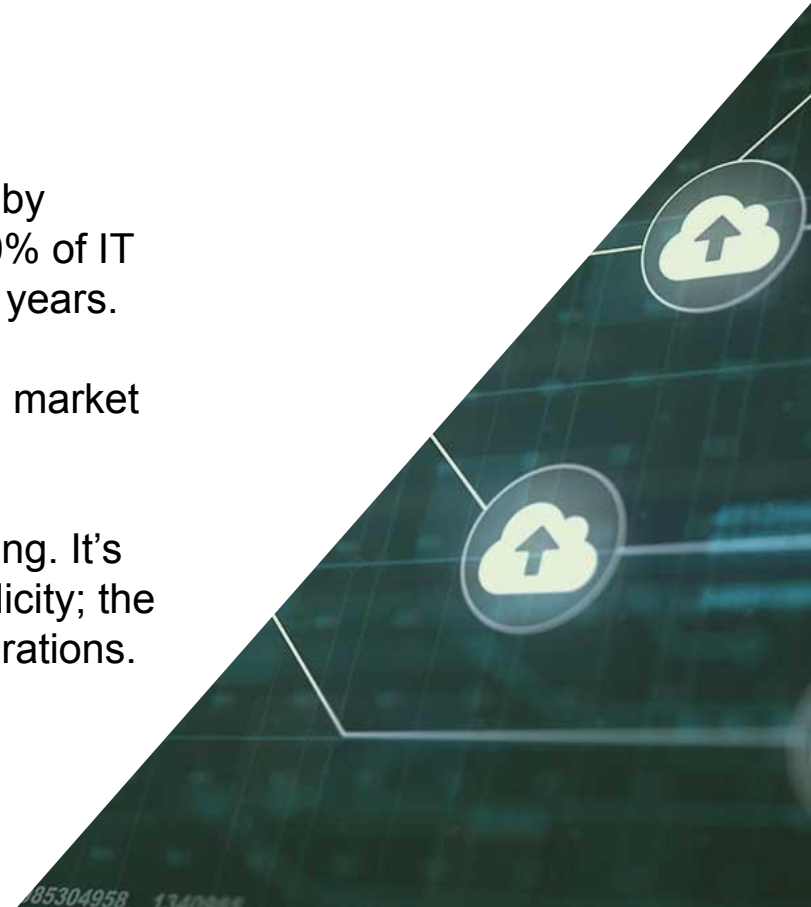
## Keynote: Enrich Big Data on the Cloud

**Sergio Rodriguez de Guzmán**

CTO PUE

# The Age of Cloud Computing

- The cloud has taken the world of computing by storm, and experts predict that more than 50% of IT will be in the cloud within the next five to ten years.
- It has been reported that by 2020, the global market for cloud equipment will reach $90.1 billion.
- With the passage of time cloud is also evolving. It's no longer enough to deliver speed and simplicity; the cloud must also support mission- critical operations.

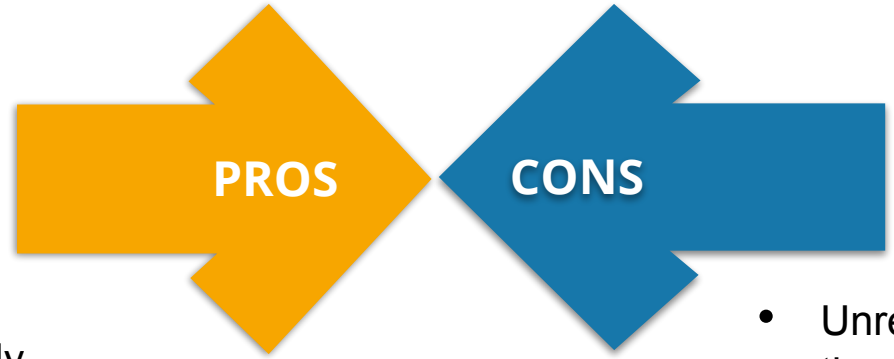


## The Public Cloud

- The most recognizable model of cloud computing to many consumers is the public cloud model, under which cloud services are provided in a virtualized environment, constructed using pooled shared physical resources, and accessible over a public network such as the internet.

- A third-party provider offers shared and scalable resources, such as applications and storage servers, which users may access over the Internet.

# Public Cloud



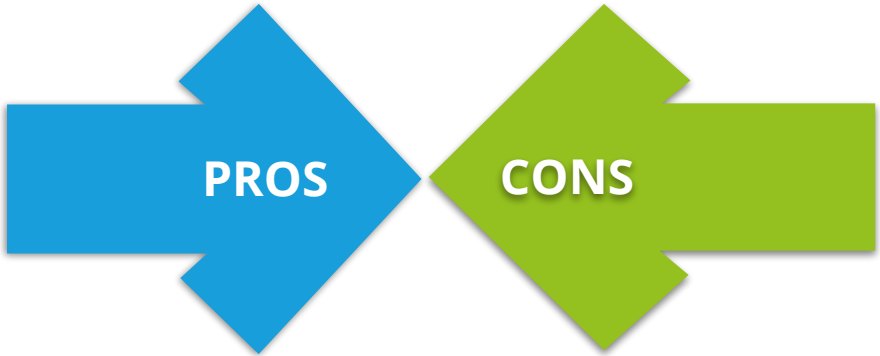

- Simple
- User friendly
- Flexibility
- High Speed
- Time Saving
- Pay-Per-Use Model
- Universal Accesibility

- Unreliability or outages at the cloud server end
- Susceptible to online hacking
- Variable broadband Internet connections could cut data off entirely

## The Private Cloud

- Private cloud is the phrase used to describe a cloud computing platform that is implemented within the corporate firewall, under the control of the IT department.
- Third-party providers may have a role in running and managing the platform, but it remains under the operational control of the enterprise.
- IDC estimates 69% of manufacturers worldwide now have more than two applications in a private cloud.

## Private Cloud




- Greater Control Over Data
- More Data Integrity
- Easy customization

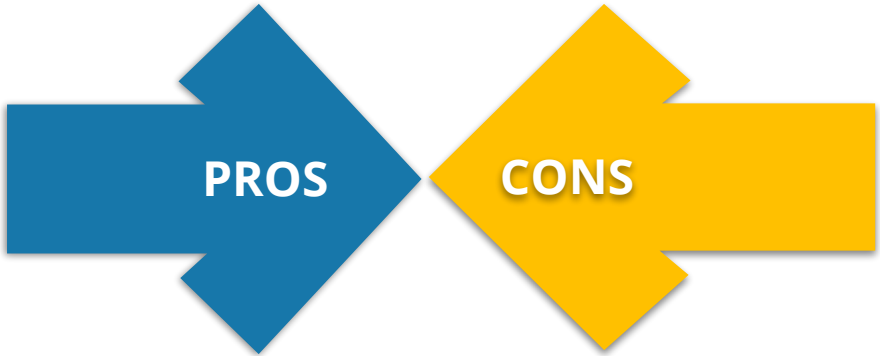
- Cost of installation and maintenance is high
- Easy scalability is not possible
- More workforce required

# The Hybrid Cloud

- The hybrid cloud is a “mix and match” approach, where the enterprise picks and chooses various elements from either the public cloud or private cloud. It is a radical approach, trying to build a cloud model that suits a specific application best.

- The hybrid cloud works well for enterprises that are capable of splitting their data into “sensitive” and “non sensitive” spheres, and big data.

# Hybrid Cloud




- Best features of private and public cloud
- Flexible according to the needs
- Most cost effective mode

- Needs regular monitoring
- Needs more maintenance
- Costliest model due to too much of customization

## Facts

### 01

The hybrid cloud is nevertheless the fastest growing cloud model compared to other models.

### 02

About **82%** of enterprises now adopt a multi-cloud strategy, while only **10%** use a single public cloud and another **5%** use a single private cloud.

### 03

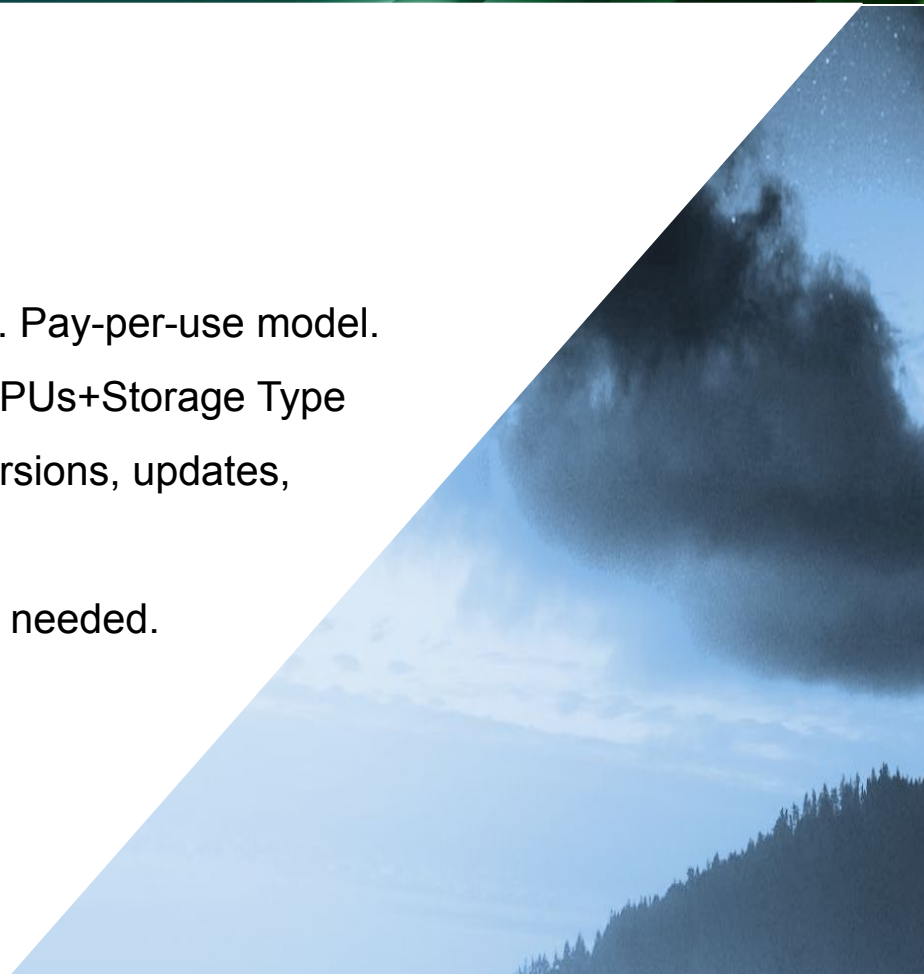
Among those enterprises who adopt a multi-cloud strategy, **14%** of enterprises use multiple private clouds, **13%** use multiple public clouds, and **55%** use hybrid clouds.

## Big Data On-Premise

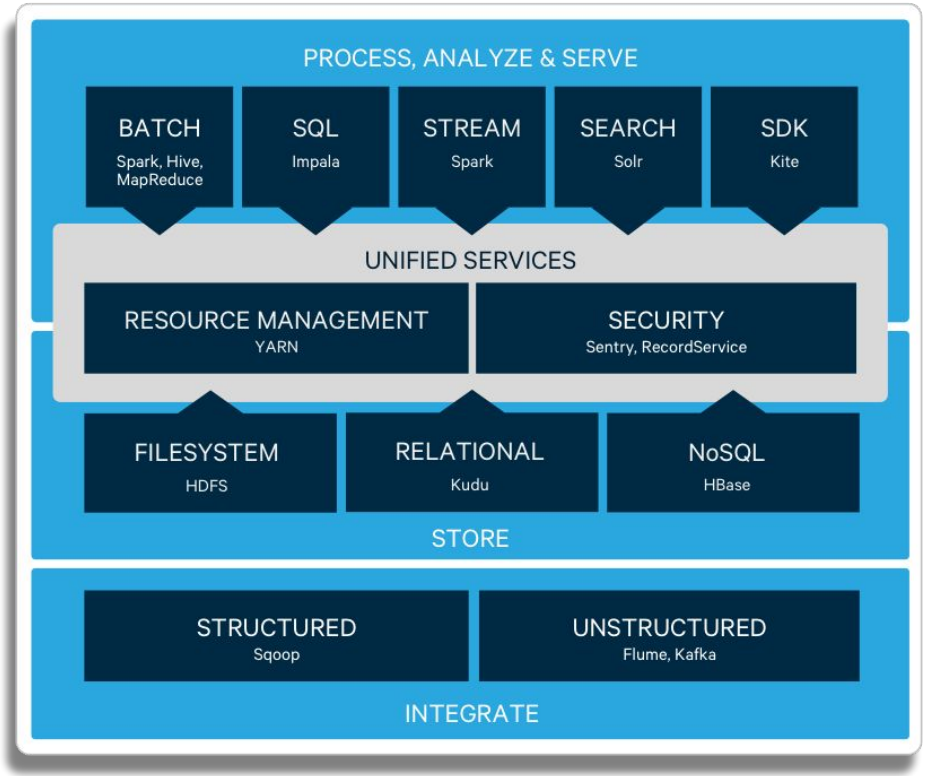
- Low hardware utilization
- Lack of multi-tenancy support
- No self-serve model
- Slow onboarding new applications/users
- Low bandwidth network
- High OPEX
- Lack of big data skills and expertise

# Big Data on the Cloud

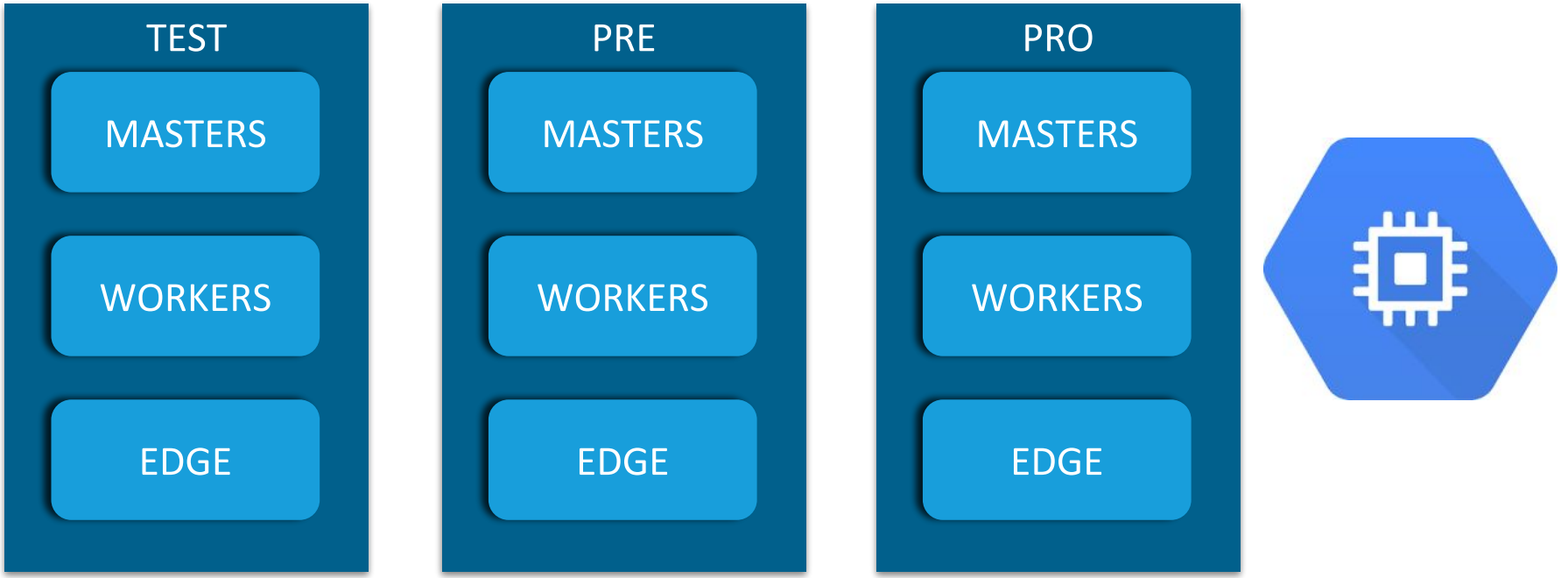
- Lower cost: no hardware investment. Pay-per-use model.
- Speed: Customize CPU+Memory+GPUs+Storage Type
- Comfort: Reduce the overhead of versions, updates, security issues…
- High level of customization: scale as needed.
- Accessibility
- Security



# Cloudera Architecture



# Enrich Infrastructure



# Enrich Infrastructure

TEST
- MASTERS
- WORKERS
- EDGE

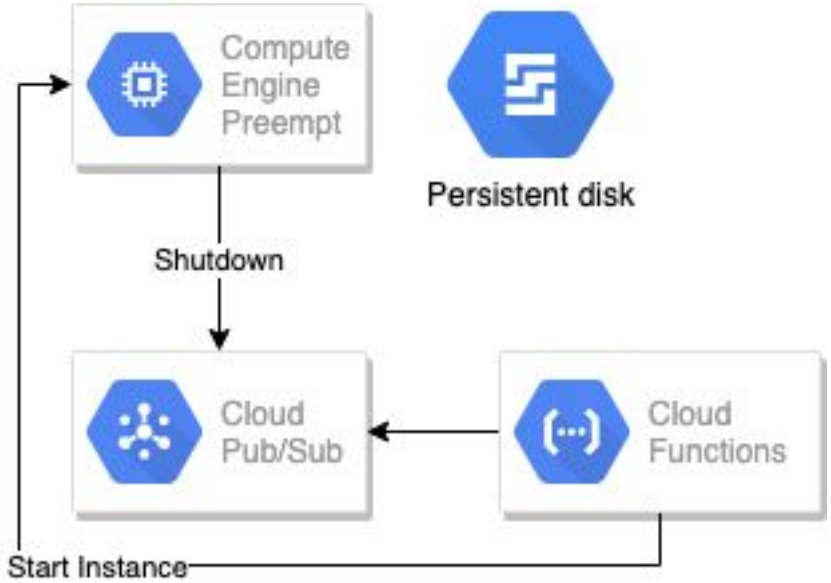
PRE
- MASTERS
- WORKERS
- EDGE

Primary WORKERS

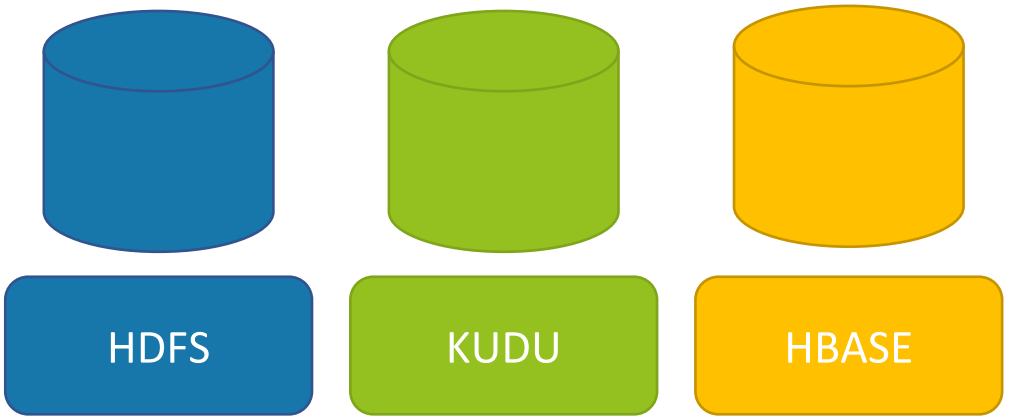
Secondary WORKERS

Preemptible Instances

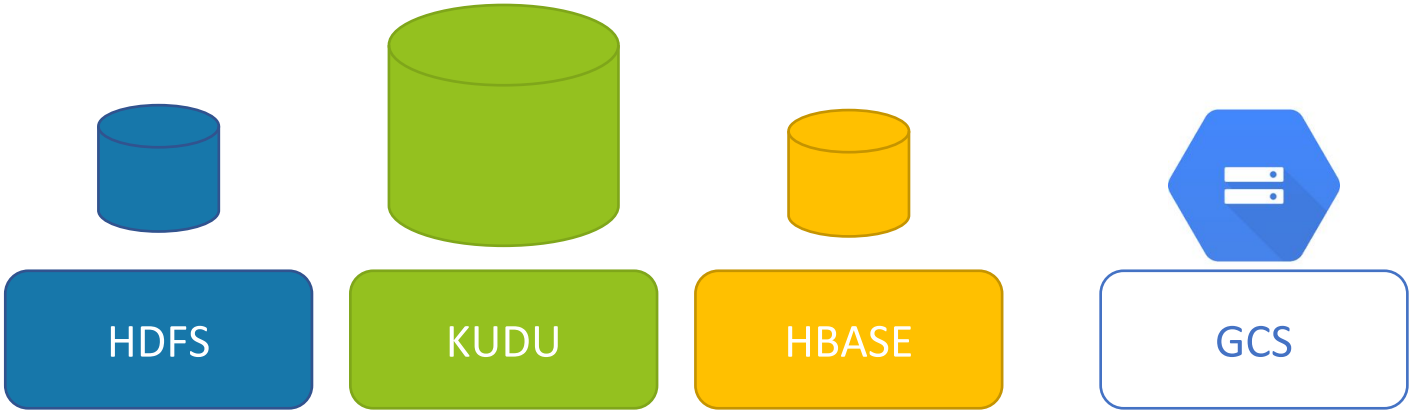
# Enrich Infrastructure



# Enrich Storage



# Enrich Storage



# HDFS & GCS

HDFS

- Replica x 3
- Magnetic or SSD (network storage)
- Dedicated storage
- Virtual or dedicated disks
- Encrypted and encrypted at rest
- Zone dependant

- No replicas
- Network storage
- No encryption
- Multi-Regional/Regional
- Nearline
- Coldline

# HDFS & KUDU & GCS

HDFS

- Sensitive data

KUDU

- Fast data
- Frequent CRUD ops

- Spark Apps
- Raw/Landing data
- Historical data
- Backups

# GDPR Vs. Data Localization

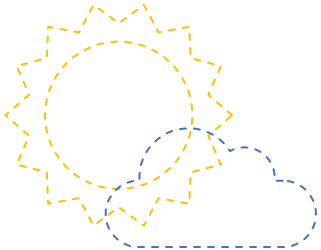
- Data localization laws restrict the storage of personal data to within the borders of a particular country or region. A frequent misunderstanding about GDPR is that personal data must remain within the EU. This is not the case.

- Specifically, personal data can be moved outside the EU, but only if the jurisdiction in which the recipient is located provides an adequate level of data protection. However, outside the EU, multiple global data localization laws do exist, including laws in Canada, China, Australia, and Russia.

# Data Localization vs. Public Cloud

The challenge for organizations is to ensure they meet local data protection regulations where they exist, yet retain the flexibility to fully use their public cloud infrastructure in regions where strict data localization rules don’t apply.

# Public cloud vs. GDPR



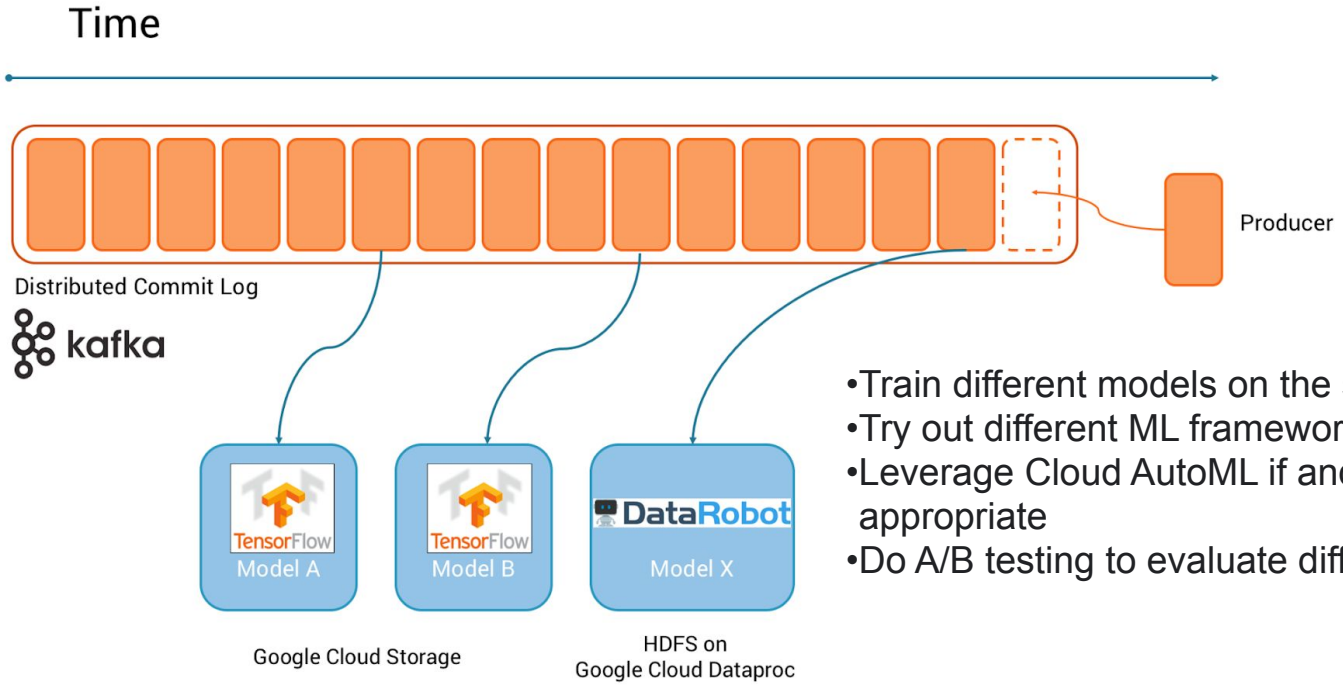
In order to support GDPR compliance in the public cloud, users need to know in near real time where their data is being stored, moved, and processed. They need to be able to configure and enforce rules that ensure that their business data is only moved to, processed, and stored in regions the European Commission has recognized as having adequate levels of data protection.

# Enrich Process, Analyze & Serve

# Enrich Process, Analyze & Serve

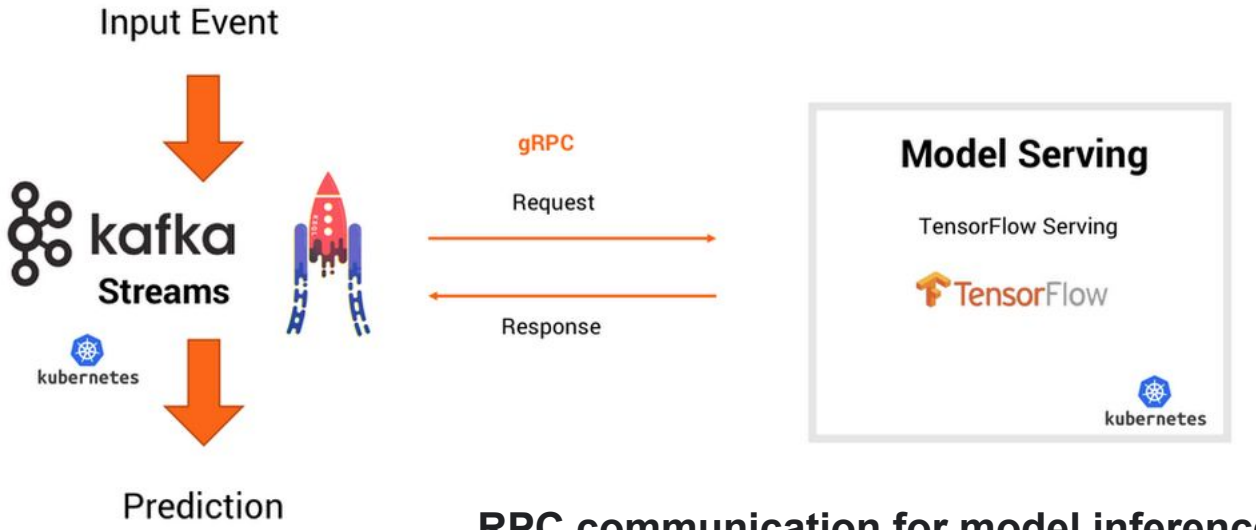
**Model training with Cloud ML Engine and TensorFlow**

# Enrich Process, Analyze & Serve



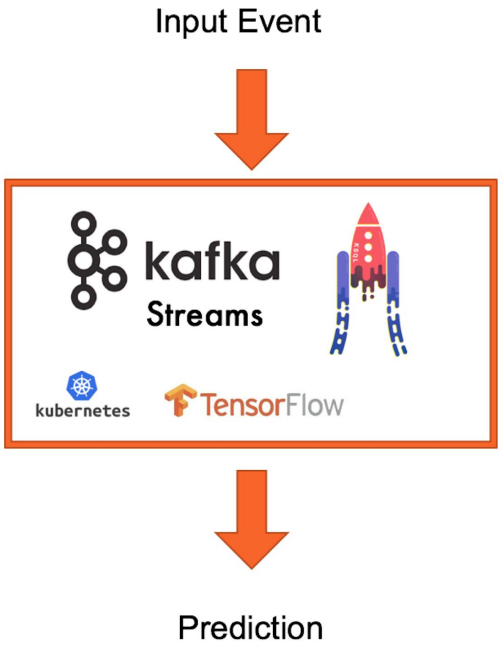

- Train different models on the same data
- Try out different ML frameworks
- Leverage Cloud AutoML if and where appropriate
- Do A/B testing to evaluate different models

# Enrich Process, Analyze & Serve




**RPC communication for model inference on your model server**

# Enrich Process, Analyze & Serve




**Integrate model inference natively into your streaming application**

CLOUD
BIG DATA

innovation
hub sessions

# ¡Gracias!

#PUEihubSessions

consulting@pue.es

BCN: 93 206 02 49

MAD: 91 162 06 69

Organiza:

pue

Participan:

cloudera

Google Cloud

confluent

MESOSPHERE

redislabs
HOME OF REDIS