



How to Build Data Science Pipelines with OpenShift using Ceph, Kafka and Knative

Guillaume Moutier
Sr Principal Technical Evangelist



About me

- Guillaume Moutier
- Sr Principal Technical Evangelist in the Cloud Storage and Data Services BU since 08/2019
- Former CTO of Laval University, Quebec, Canada
- Working on Data Science platforms, Data Engineering, Data..., Data..., Data...

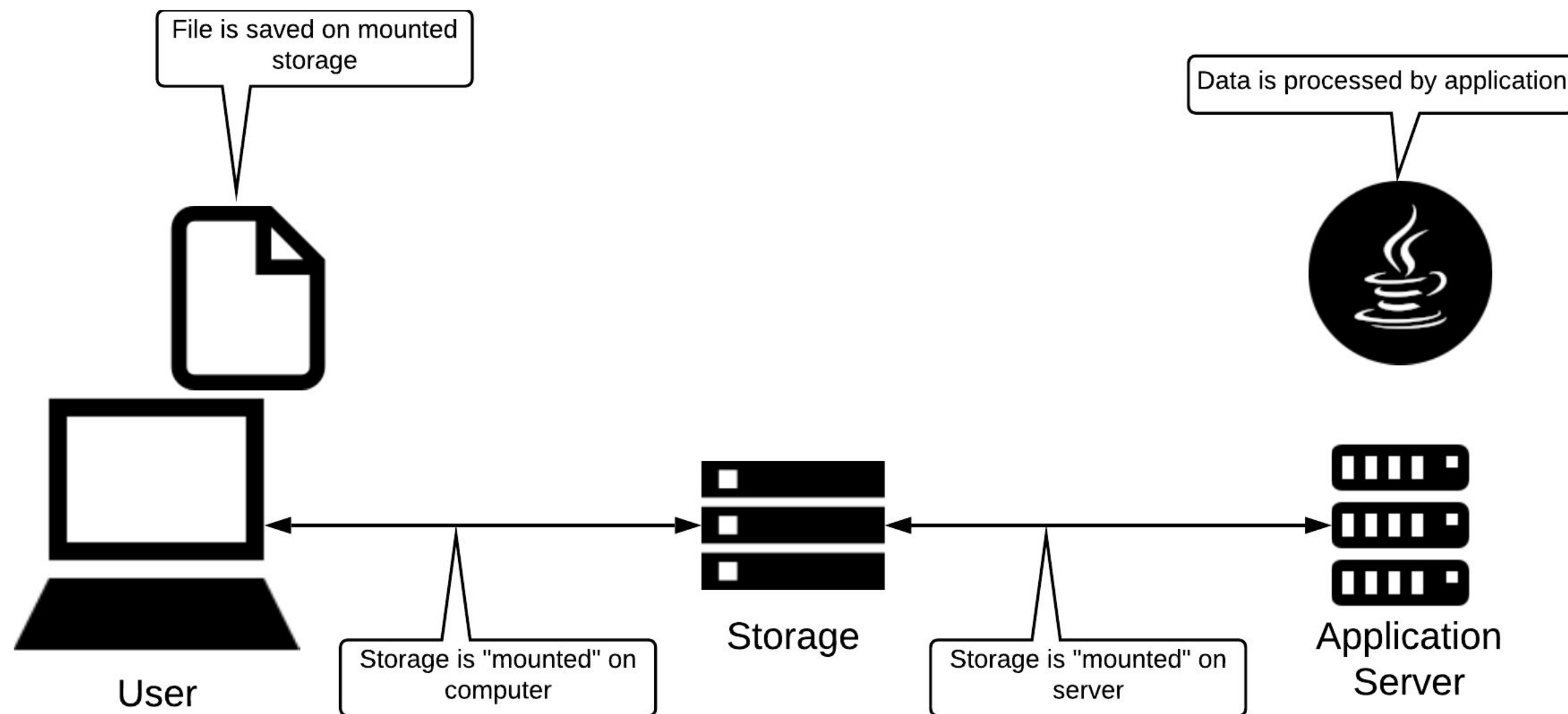
The Cloud Native way of doing things!

(Opinionated) Characteristics for a Cloud Native data platform:

1. **Agility and Elasticity**: as tools, frameworks and datasets evolve constantly and rapidly, you must be able to act accordingly.
2. **Cloud standards**: avoid vendor lock-in with proprietary tools and formats, and embrace widely recognized open-source protocols and standards.
3. **Hybrid cloud architecture**: your architecture must run anywhere without any change (some configs may be adapted, but not the architecture itself).
4. **Automation**: embrace the devops philosophy. Everything must be automated and code-based.
5. **Separate Compute from Storage**: take advantage of the rich computing ecosystem against central storage.

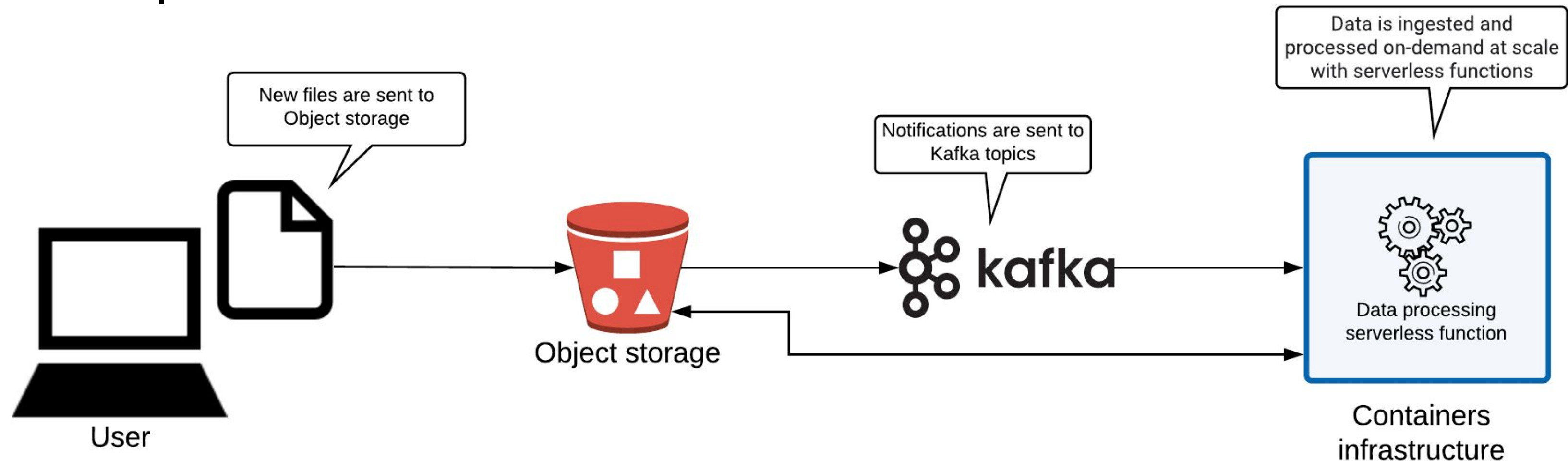
Outcomes: speed, efficiency, adaptability

Legacy data pipeline architecture (still standard?)



Everything tightly coupled and not easily scalable

Example of a Cloud Native architecture pattern

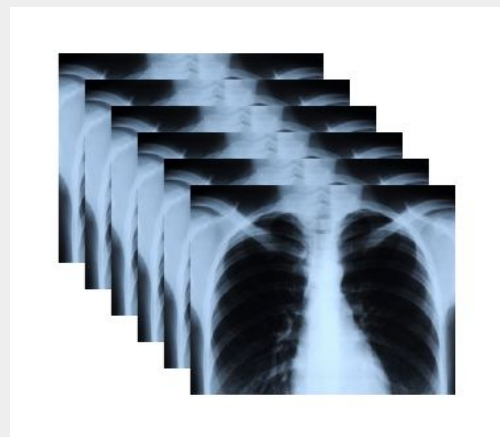


Everything disconnected and automatically scalable!

Demo: Automated XRay analysis

Introduction of the Use-Case

Pneumonia detection from chest x-rays using an automated data pipeline



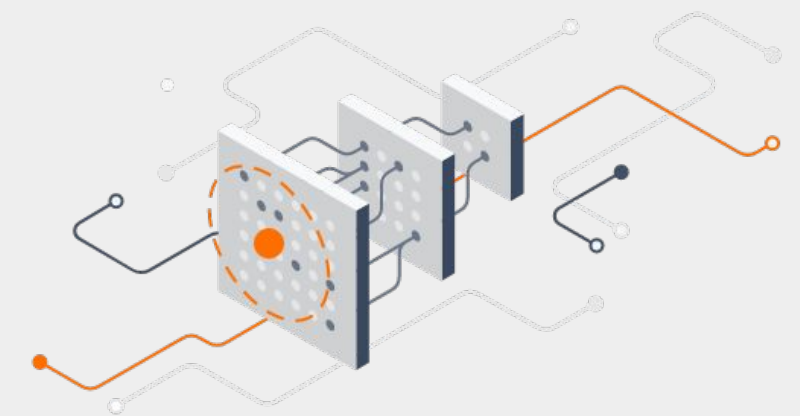
Pneumonia



Normal



OPEN DATA HUB
AI Platform powered by Open Source

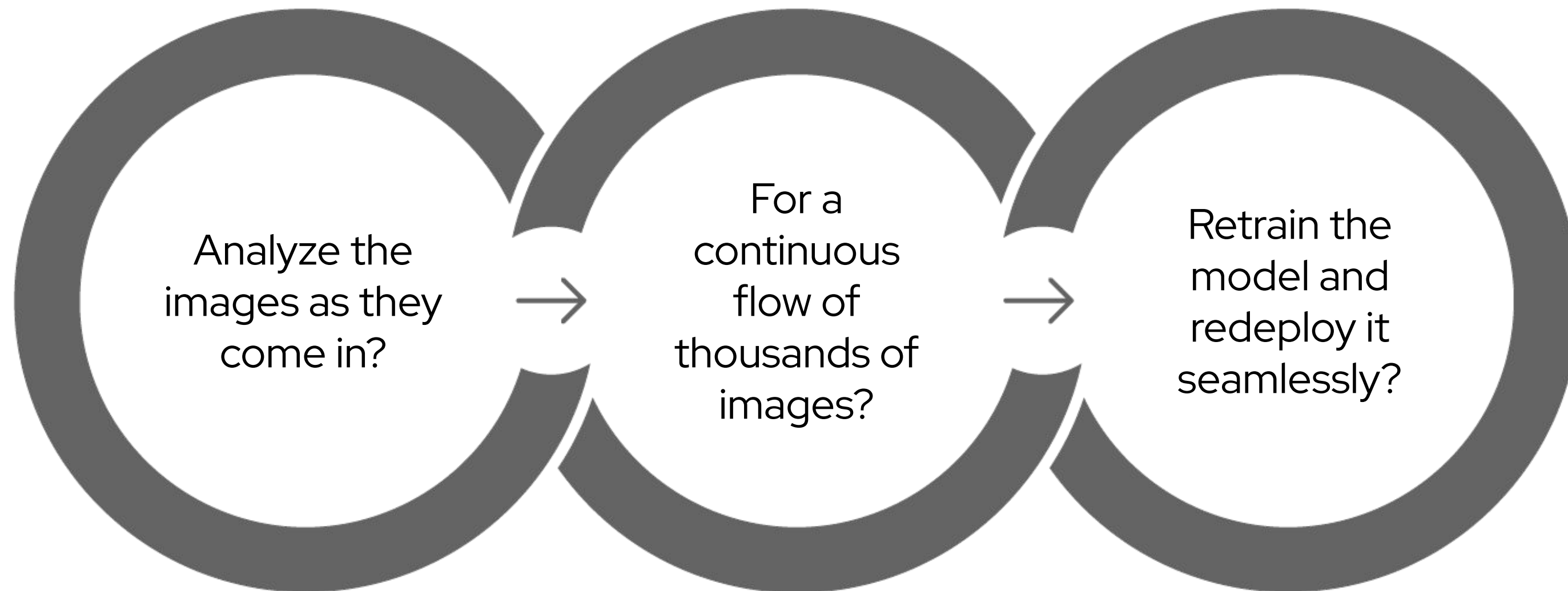


The problem: Imagine we have X-Ray images to review.

An AI/ML model can help!

Medical images analysis has to scale, let's automate it!

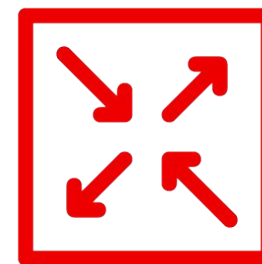
Now we have a model!
But how can we efficiently...

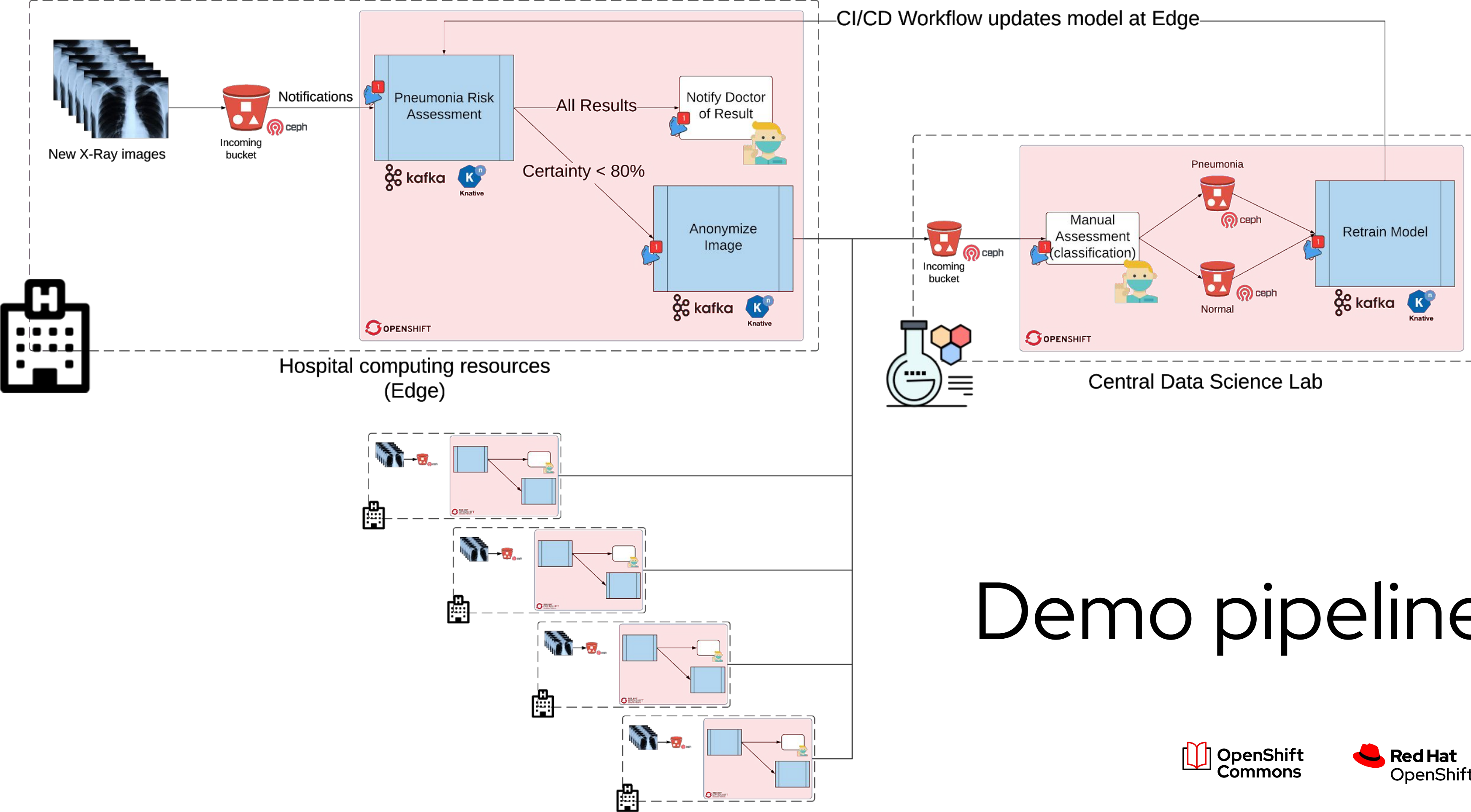


At various locations simultaneously?

The answer: Cloud native architecture and patterns.

- Bucket notifications with OpenShift Container Storage
- Kafka Topics with AMQ Streams
- KNative Eventing and Serving with OpenShift Serverless





Contact Me!

gmoutier@redhat.com

Learn the code, source available:

<https://github.com/guimou/datapipelines>


Visit our websites

Learn more about our AI/ML capabilities, and see success stories from existing customers.

openshift.com/ai-ml, opendatahub.io

Watch our videos

Visit our [YouTube channel](#) to discover a wide range of videos answering all your AI/ML questions.

 linkedin.com/company/red-hat/

 facebook.com/openshift

 [.youtube.com/OpenShift](https://youtube.com/OpenShift)

 twitter.com/openshift

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

 linkedin.com/company/red-hat/

 facebook.com/openshift

 [.youtube.com/OpenShift](https://youtube.com/OpenShift)

 twitter.com/openshift