# Cluster Serving:

## Distributed and Automated Model Inference on Big Data Streaming Frameworks

Authors: Jiaming Song, Dongjie Shi, Qiyuan Gong, Lei Xia, Jason Dai

intel

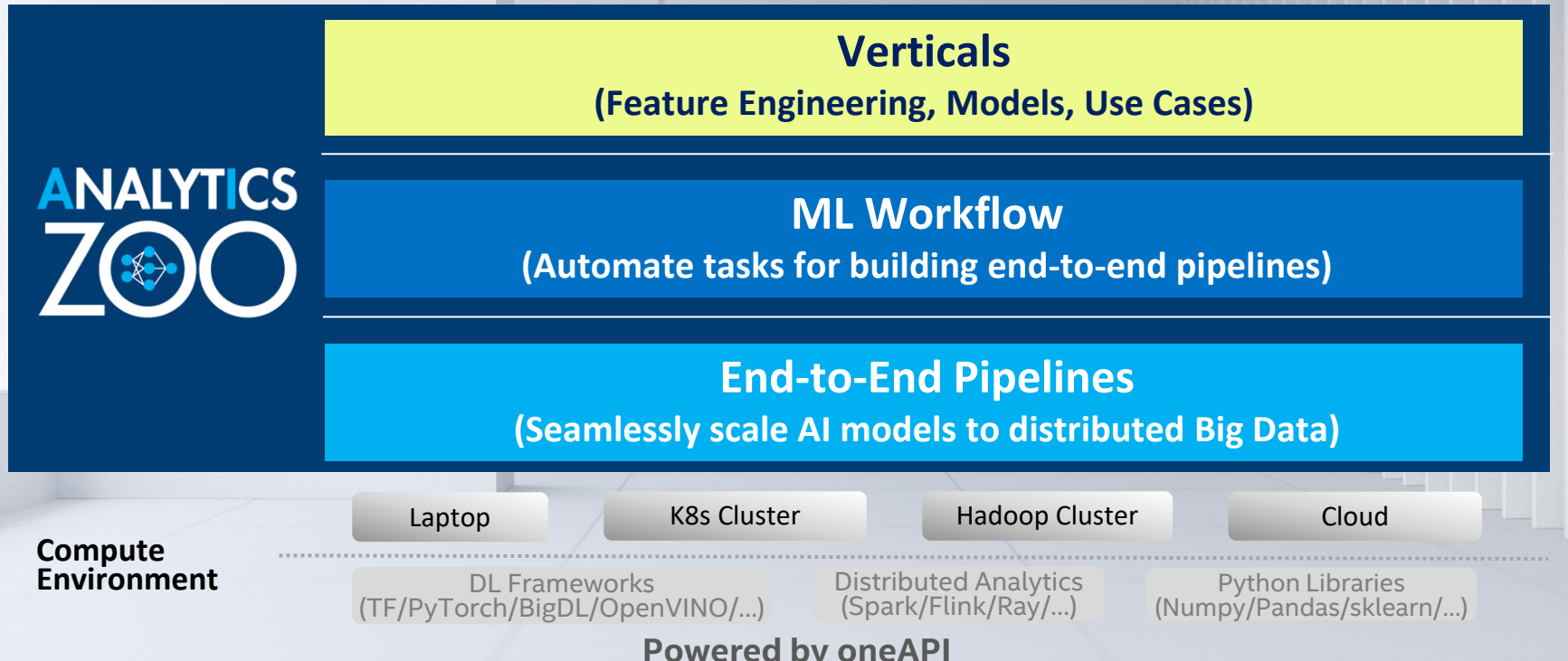
# Agenda

- Analytics Zoo: Software Platform for Big Data AI
- Cluster Serving on Analytics Zoo
- Summary

# Agenda

- **Analytics Zoo: Software Platform for Big Data AI**
- **Cluster Serving on Analytics Zoo**
- **Summary**

# Analytics Zoo: Software Platform for Big Data AI

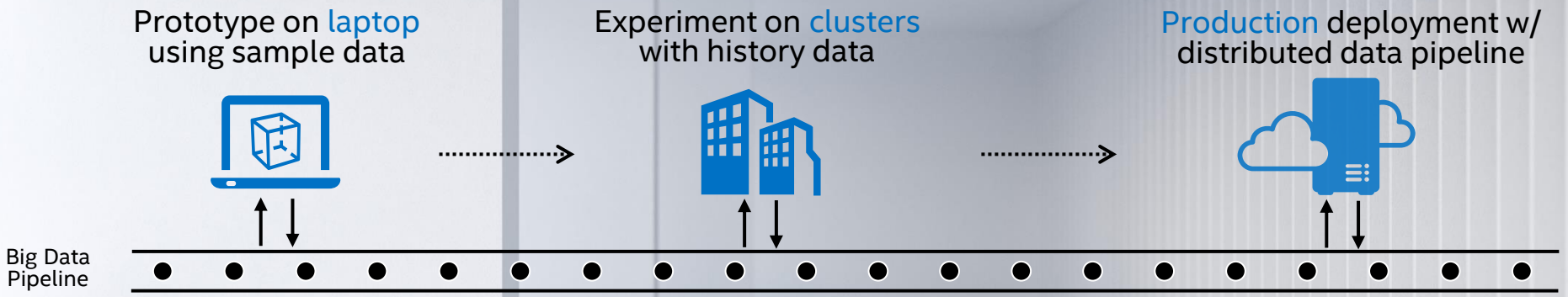



https://github.com/intel-analytics/analytics-zoo

intel

# End-to-End Big Data Analytics and AI

## Seamless Scaling from Laptop to Distributed Big Data



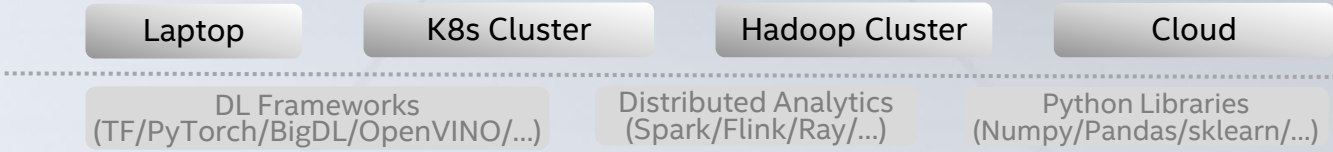

- Easily prototype end-to-end pipelines that apply AI models to big data
- "Zero" code change from laptop to distributed cluster
- Seamlessly deployed on production Hadoop/K8s clusters
- Automate the process of applying machine learning to big data

intel

# Analytics Zoo: Open Source Platform for Big Data AI

## Scaling End-to-End AI to Distributed Big Data

| | |
|---|---|
| **PPML** | Privacy Preserving Data Analytics & ML on SGX |
| **Zouwu** | Scalable time series analysis pipeline w/ AutoML |
| **RayOnSpark** | Run Ray programs directly on Big Data platform |
| **Cluster Serving** | Distributed real-time model serving on Flink |
| **Orca** | Seamlessly scale out TF & PyTorch on Spark & Ray |




**Powered by oneAPI**

https://github.com/intel-analytics/analytics-zoo

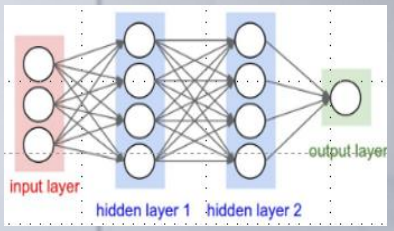
intel

# Agenda

- Analytics Zoo: Software Platform for Big Data AI
- **Cluster Serving on Analytics Zoo**
- Summary

# Serving

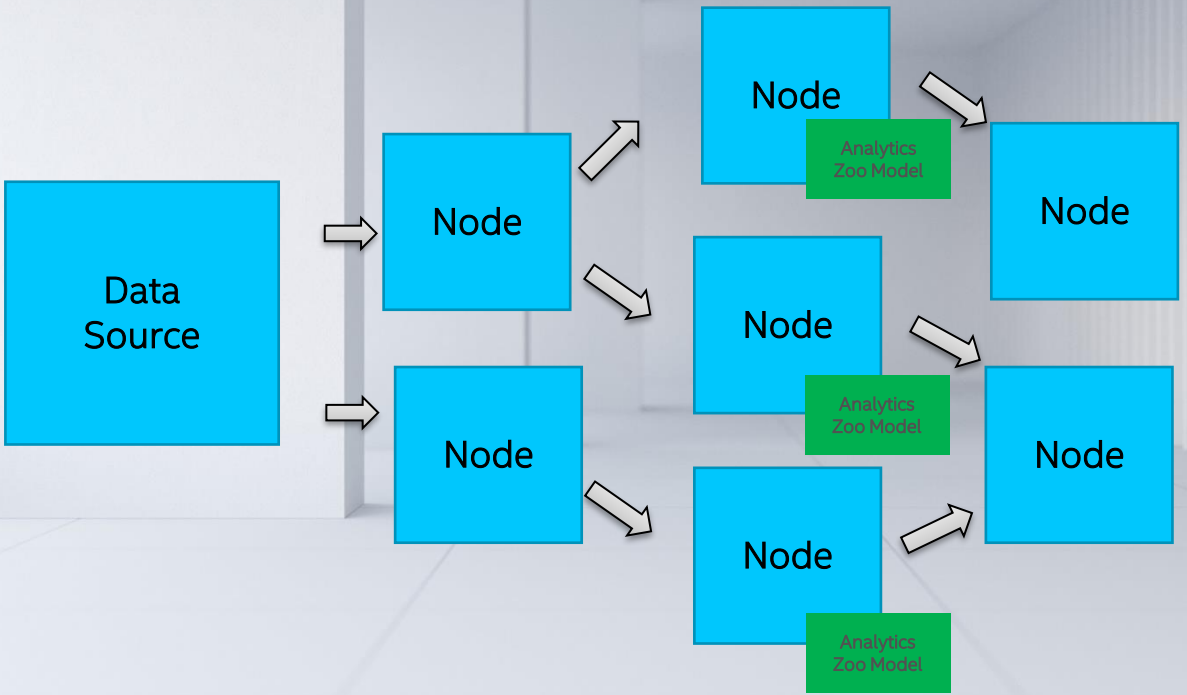
Use trained model to serve end-to-end ML pipeline




model

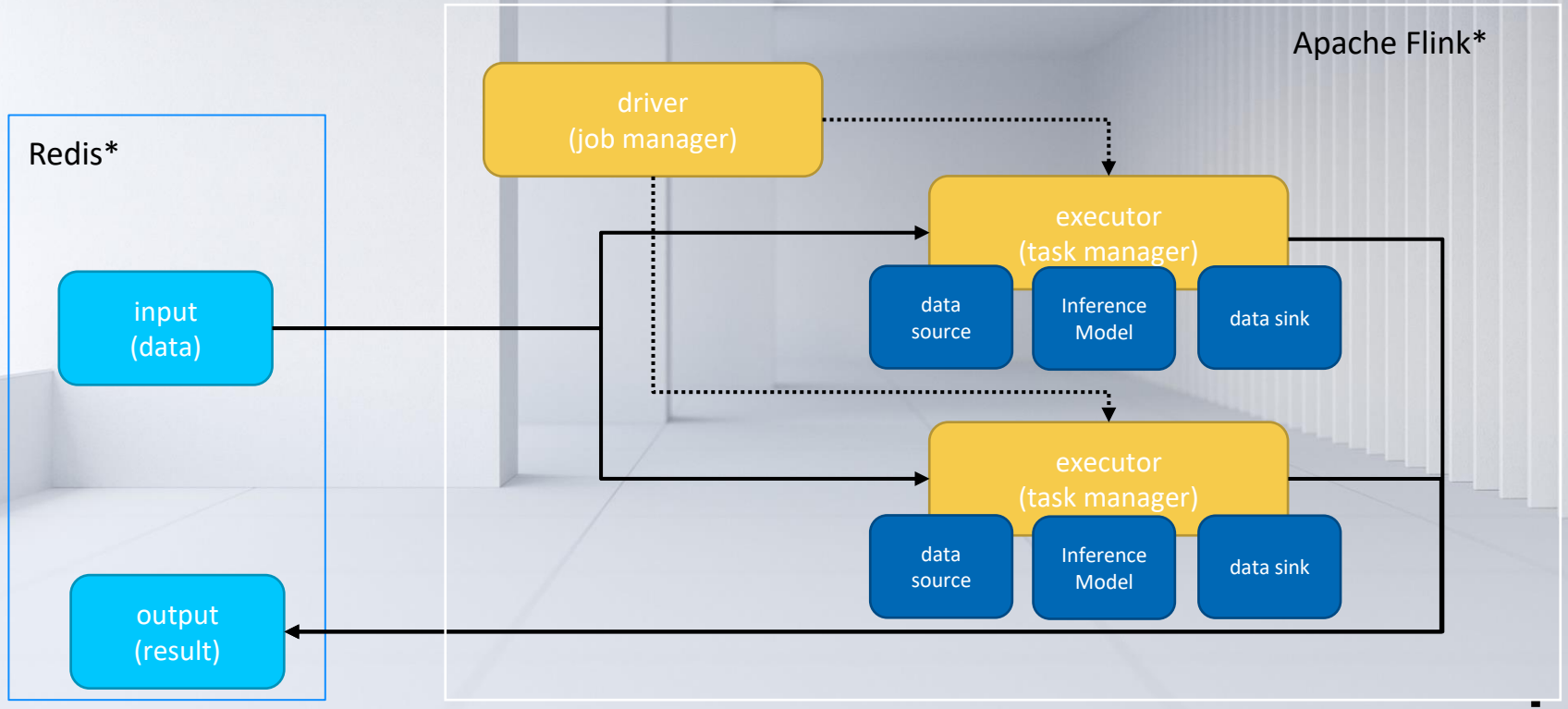
Input Data → Preprocessing → Inference → Postprocessing → Result

# Distributed Model Serving



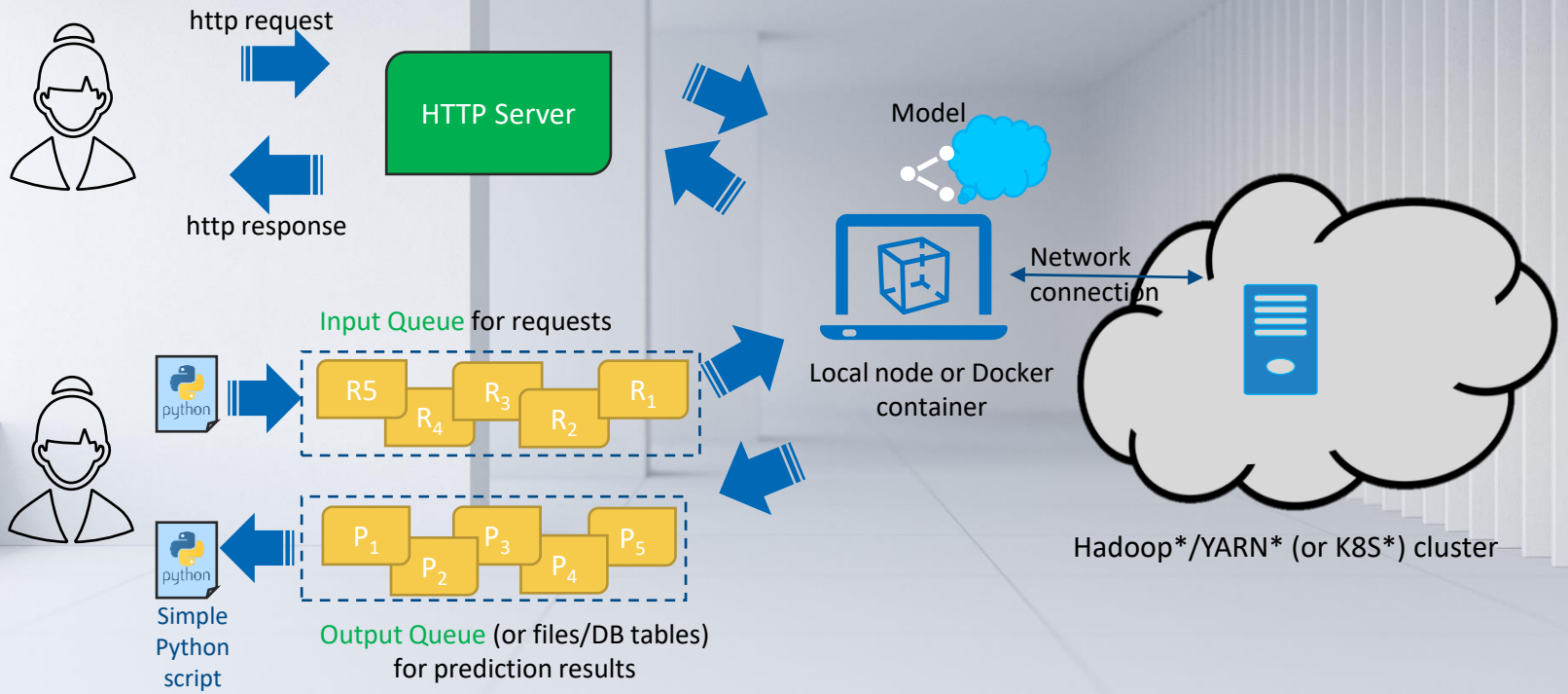

intel

# Architecture of Cluster Serving on Analytics Zoo, Flink, Redis




intel

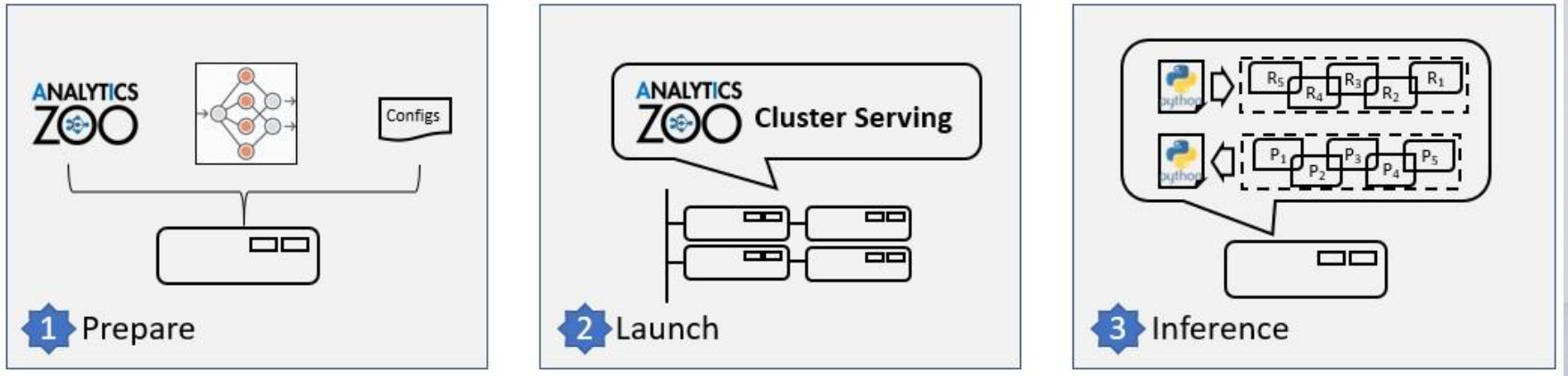
# Data pipeline User Perspective




intel

# Cluster Serving Workflow Overview

1. Install and prepare Cluster Serving environment on a local node
2. Launch the Cluster Serving service
3. Distributed, real-time (streaming) inference




intel

# Very Quick Start

## Start docker container

```
#docker run -itd --name cluster-serving --net=host intelanalytics/zoo-cluster-serving:0.7.0
```

## Log into container

```
#docker exec -it cluster-serving bash
```

## Start Serving

```
#cluster-serving-start
```

https://github.com/intel-analytics/analytics-zoo/blob/master/docs/docs/ClusterServingGuide/ProgrammingGuide.md

intel

# API Introductions

## http sync API

- data are represented by json format
- call http post method to enqueue your data into pipeline
- http API is compatible with TFServing*

## pub-sub python sync/async API

- data are represented by ndarray
- call python method to enqueue your data into pipeline

# API Examples

HTTP service

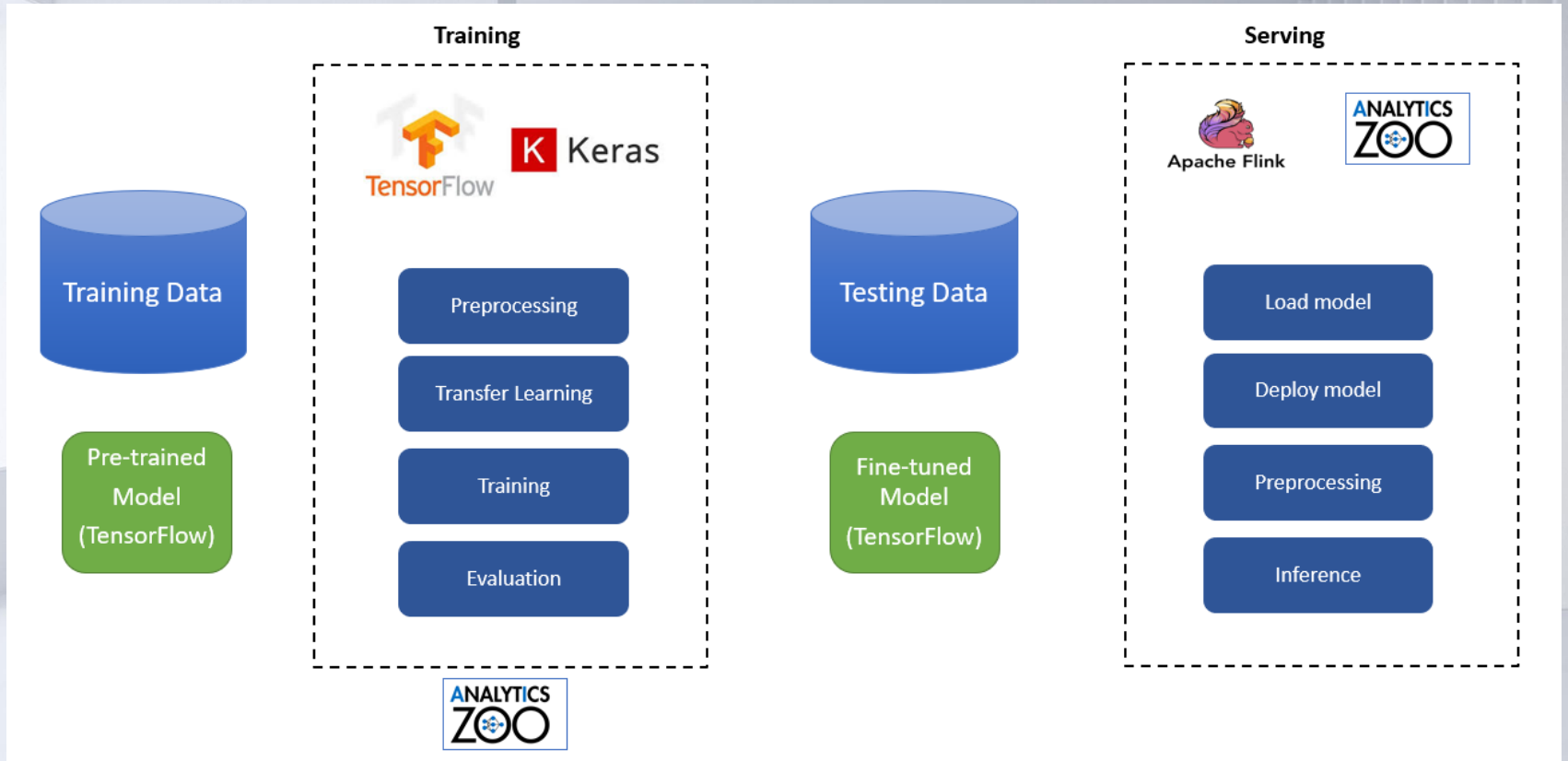
```
curl -d \
 '{
  "instances" : [ {
    "intScalar" : 12345,
    "floatScalar" : 3.14159,
    "stringScalar" : "hello, world. hello, arrow.",
    "intTensor" : [ 7756, 9549, 1094, 9808, 4959, 3831, 3926, 6578, 1870, 1741 ],
    "floatTensor" : [ 0.6804766, 0.30136853, 0.17394465, 0.44770062, 0.20275897, 0.32762378, 0.45966738, 0.30405
    "stringTensor" : [ "come", "on", "united" ],
    "intTensor2" : [ [ 1, 2 ], [ 3, 4 ], [ 5, 6 ] ],
    "floatTensor2" : [ [ [ 0.2, 0.3 ], [ 0.5, 0.6 ] ], [ [ 0.2, 0.3 ], [ 0.5, 0.6 ] ] ],
    "stringTensor2" : [ [ [ [ "come", "on", "united" ], [ "come", "on", "united" ], [ "come", "on", "united" ],
  }, {
    "intScalar" : 12345,
    "floatScalar" : 3.14159,
    "stringScalar" : "hello, world. hello, arrow.",
    "intTensor" : [ 7756, 9549, 1094, 9808, 4959, 3831, 3926, 6578, 1870, 1741 ],
    "floatTensor" : [ 0.6804766, 0.30136853, 0.17394465, 0.44770062, 0.20275897, 0.32762378, 0.45966738, 0.30405
    "stringTensor" : [ "come", "on", "united" ],
    "intTensor2" : [ [ 1, 2 ], [ 3, 4 ], [ 5, 6 ] ],
    "floatTensor2" : [ [ [ 0.2, 0.3 ], [ 0.5, 0.6 ] ], [ [ 0.2, 0.3 ], [ 0.5, 0.6 ] ] ],
    "stringTensor2" : [ [ [ [ "come", "on", "united" ], [ "come", "on", "united" ], [ "come", "on", "united" ],
  } ]
}' \
-X POST http://host:port/predict
```

```
from zoo.serving.client import InputQueue
import numpy as np
input_api = InputQueue()
t1 = np.array([1,2])
t2 = np.array([[1,2], [3,4]])
input_api.enqueue('my-instance', img={"path": 'path/to/image'}, tensor1=t1, tensor2=t2)
```

Python pub-sub

intel

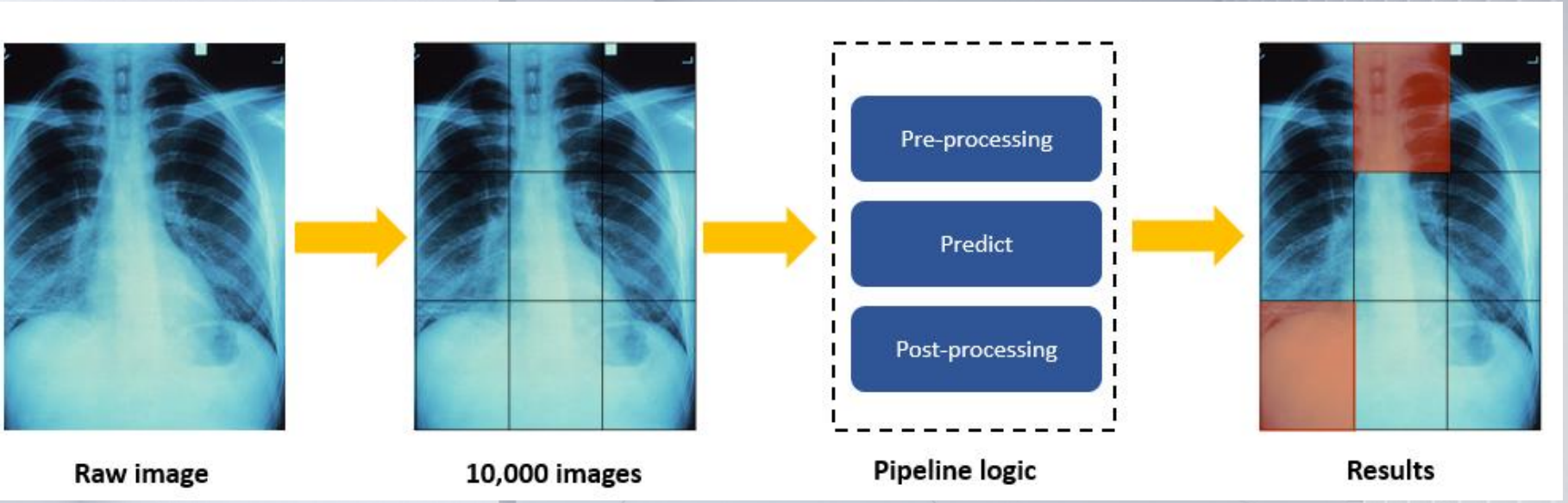
# Garbage classification on Alibaba Tianchi Competition




*Other names and brands may be claimed as the property of others.

intel

# Medical Imaging Analysis




**Bottleneck:**
**Preprocessing, inference, 10k images, up to 1-2 hours per large piece**

**scale-out the data to accelerate**

intel

# Agenda

- Analytics Zoo: Software Platform for Big Data AI
- Cluster Serving on Analytics Zoo
- **Summary**

# Advantages of Analytics Zoo Cluster Serving

## Ease of Deployment

**One container with all dependencies & leverage existed YARN/K8S cluster**

## Wide Range Deep Learning model support

**Tensorflow*, Caffe*, OpenVINO*, Pytorch*, BigDL***

## Low Latency

**Continuous Streaming pipeline is supported by Apache Flink***

## High Throughput & Scalability

**Optimization of multithread control, and could easily scale out to clusters**

# Summary

**Analytics Zoo: Software Platform for Big Data AI**

- E2E Big Data & AI pipeline (distributed TF/PyTorch/Keras/BigDL/OpenVINO/Ray on Spark)
- Vertical AI solutions (PPML, Time-Series, AutoML, etc.)

**Open Source Website**

- Project repo: https://github.com/intel-analytics/analytics-zoo
- Use cases: https://analytics-zoo.readthedocs.io/en/latest/doc/Application/powered-by.html

**Technical paper/tutorials**

- Upcoming CVPR 2021 tutorial!
- CVPR 2020 tutorial: https://jason-dai.github.io/cvpr2018/
- ACM SoCC 2019 paper: https://arxiv.org/abs/1804.05839
- AAAI 2019 tutorial: https://jason-dai.github.io/aaai2019/

# Thank You

# Notices & Disclaimers

- Performance varies by use, configuration and other factors. Learn more at www.intel.com/performanceIndex
- Performance may vary based on the specific game title and server configuration. To reference the full list of Intel Server GPU platform measurements, please refer to http://www.intel.com/content/www/us/en/benchmarks/server/graphics/IntelServerGPU
- All product plans and roadmaps are subject to change without notice.
- Intel technologies may require enabled hardware, software or service activation.
- No product or component can be absolutely secure.
- Your costs and results may vary.
- Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
- All product plans and roadmaps are subject to change without notice.
- © Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others. Intel Server GPU TCO analysis is based on internal Intel research. Pricing as of 10/01/2020. Analysis assumes standard serving pricing, GPU list pricing, and software pricing based on estimated Nvidia software license costs of $1 per year for 5 years.
- Intel Server GPU Performance may vary based on the specific game title and server configuration. To reference the full list of Intel Server GPU platform measurements, please refer to http://www.intel.com/content/www/us/en/benchmarks/server/graphics/IntelServerGPU
- Video game footage courtesy of Tencent Games and Gamestream.
- LEGO STAR WARS TITLES : © Lucasfilm Entertainment Company Ltd. or Lucasfilm Ltd. & ® or TM as indicated. All rights reserved.
- LEGO, the LEGO logo and the Minifigure are trademarks of The LEGO Group. © The LEGO Group. All rights reserved.
- “DiRT4”™ : © 2017 The Codemasters Software Company Limited ("Codemasters"). All rights reserved. "Codemasters"®, “EGO”®, the Codemasters logo, and “DiRT”® are registered trademarks owned by Codemasters. “DiRT4”™ and “RaceNet”™ are trademarks of Codemasters. All rights reserved. Under licence from International Management Group (UK) Limited. All other copyrights or trademarks are the property of their respective owners and are being used under license. Developed by Codemasters.

intel.