

ANALYTICS ZOO & CLUSTER SERVING

数据分析+AI平台技术及案例研究

AGENDA

Part 1: Intel Analytics-Zoo & Use cases

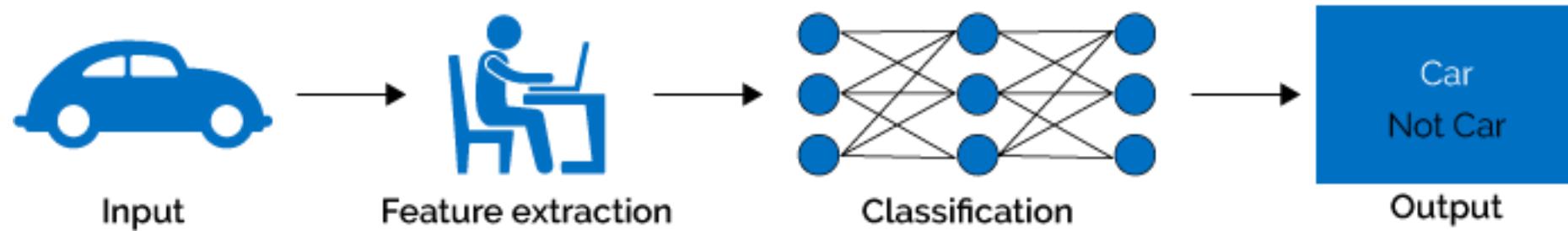
Part 2: Cluster Serving with Analytics Zoo

INTEL ANALYTICS-ZOO & USE CASES

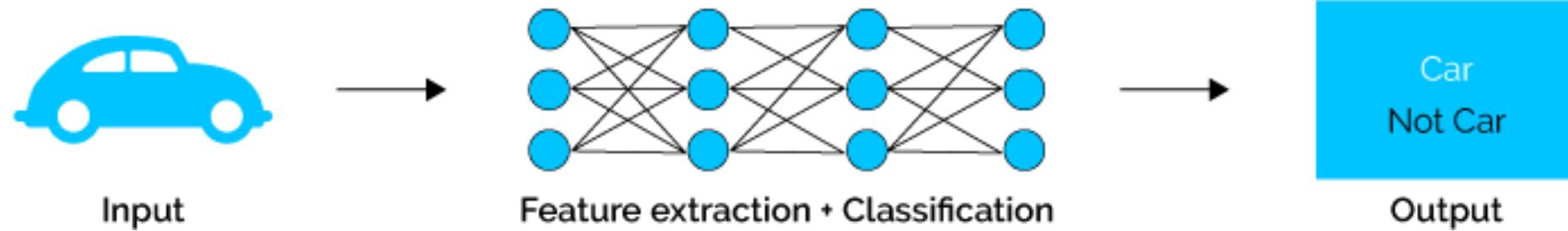


深度学习

Machine Learning

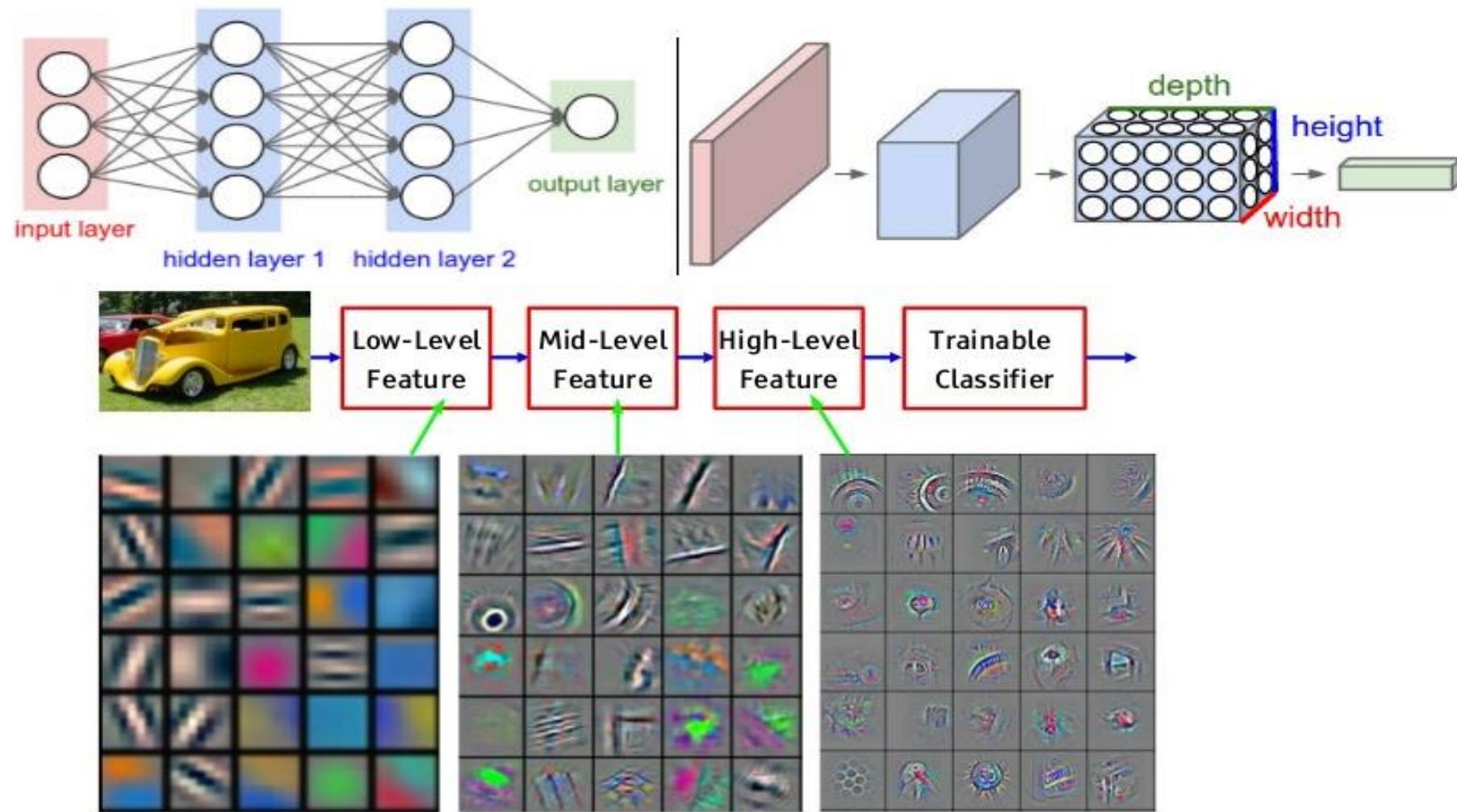


Deep Learning

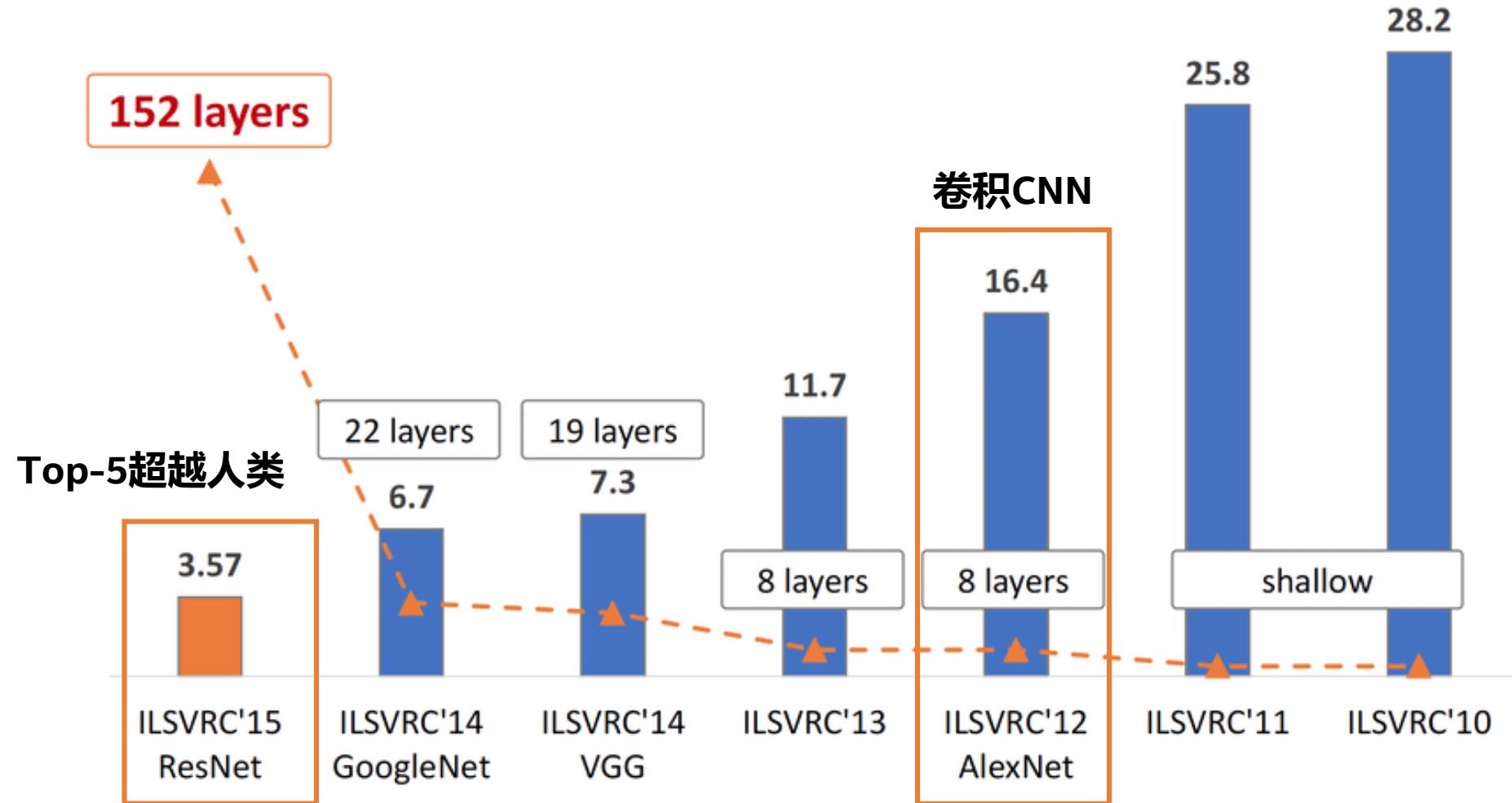


<https://www.quora.com/What-is-the-difference-between-deep-learning-and-usual-machine-learning>

深度学习CV



深度学习CV-TOP5 ERROR RATE



Kaiming etc Deep Residual Learning for Image Recognition, 2015

深度学习

2016年 Google Alpha Go 击败前世界冠军李世石

In "Nature" 27 January 2016:

"DeepMind's program AlphaGo beat Fan Hui, the European Go champion, five times out of five in tournament conditions..."

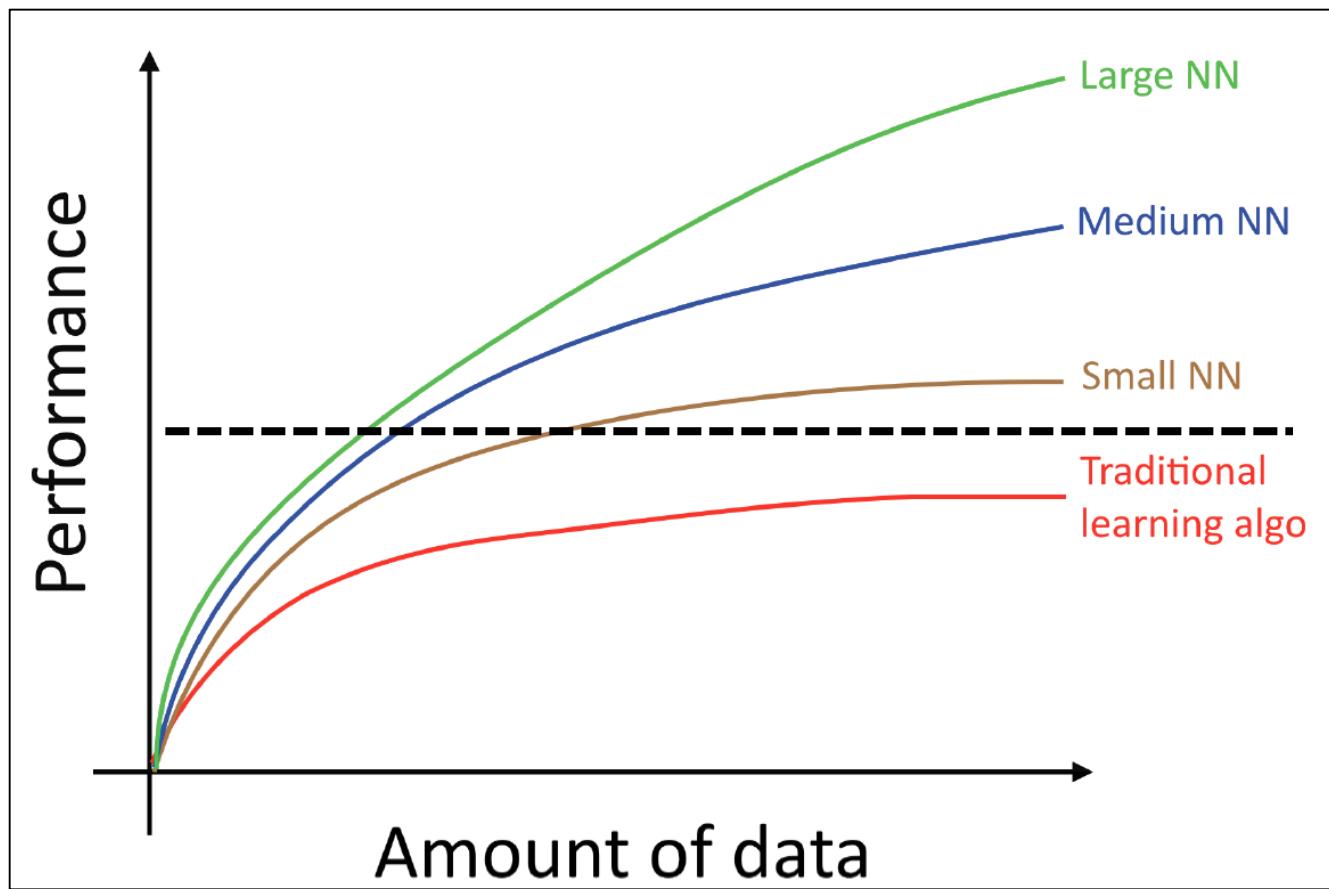
"...AlphaGo program applied deep learning in neural networks (convolutional NN) — brain-inspired programs in which connections between layers of simulated neurons are strengthened through examples and experience."

2017年 Google Alpha Go Master 击败世界冠军柯洁

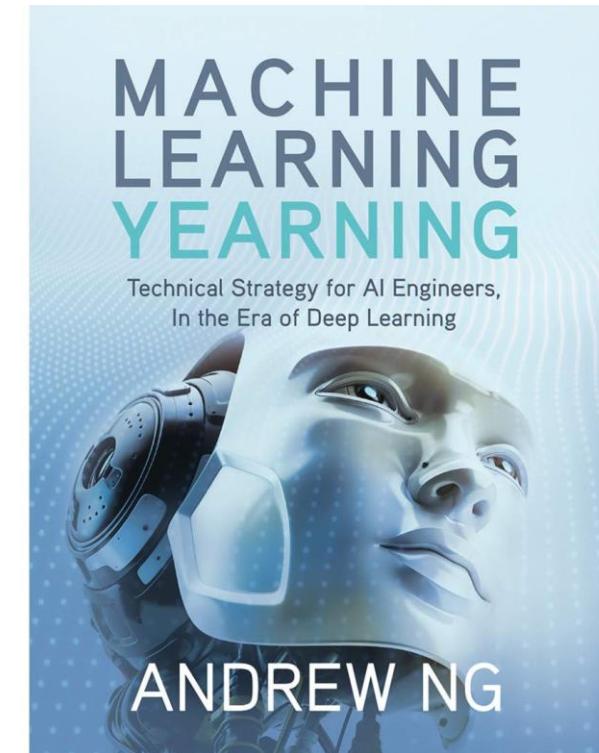


<https://www.nature.com/nature/volumes/529/issues/7587>

深度学习



“Machine Learning Yearning”,
Andrew Ng, 2016



深度学习面临的问题

但天下没有免费的午餐 (no free lunch)

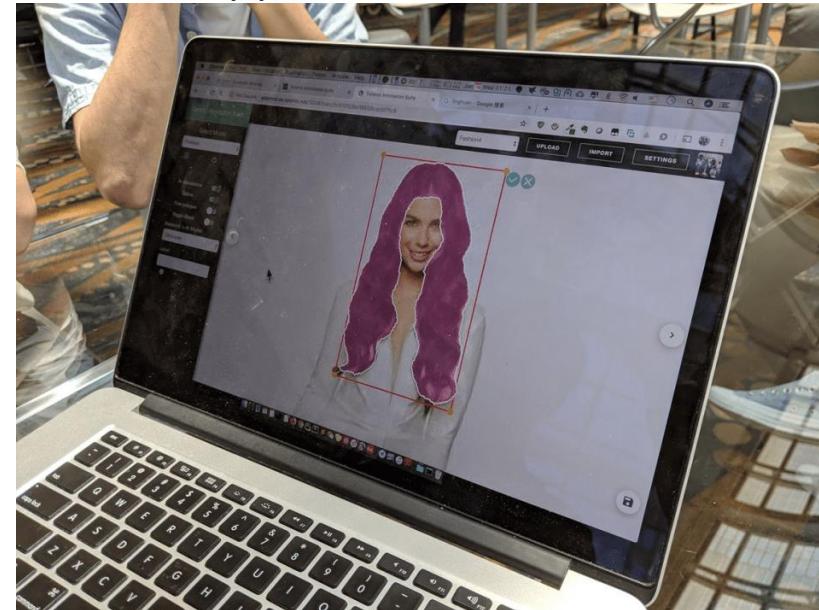
- Deep Learning需要大量**算力 (计算密集)**
- Deep Learning需要大量**数据 (data hungry)**

足够的存储和算力



IMAGENET

“人工”智能



<https://medium.com/syncedreview/data-annotation-the-billion-dollar-business-behind-ai-breakthroughs-d929b0a50d23>

深度学习

2016年 Google Alpha Go 击败李世石

- 现场的Alpha Go所使用的计算资源
 - **48 CPU, 8 GPU ≈ 24 Servers**

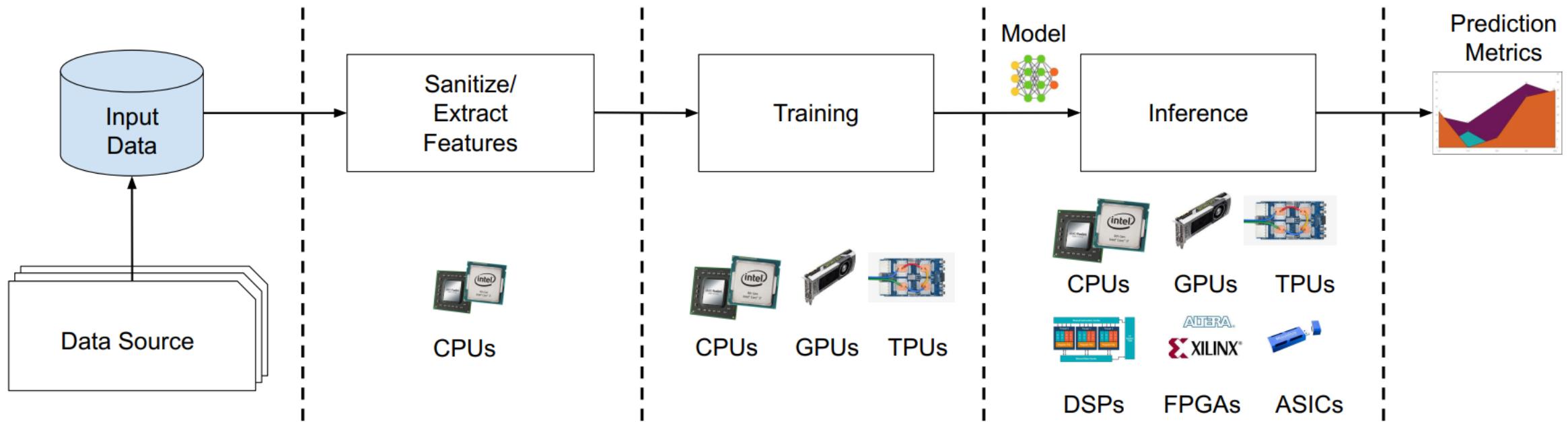


- 其实还有一个Distributed Alpha Go
 - **1202 CPU, 176 GPU ≈ 601 Servers**



<https://newsroom.intel.com/editorials/re-architecting-data-center-intel-xeon-processor-scalable-family/#gs.hi35lo>

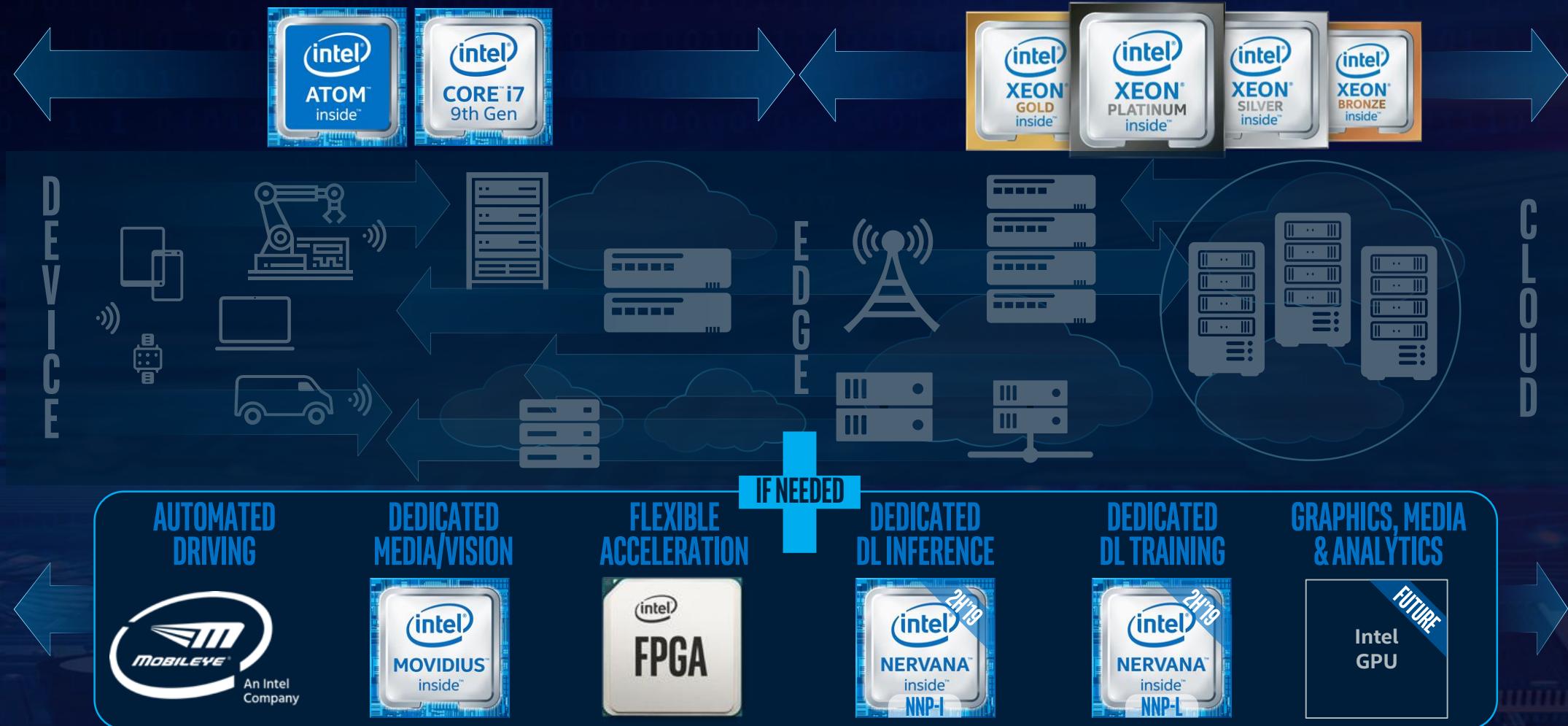
深度学习



<https://mlperf.org/>

VJ Reddi, 2019, MLPerf Inference Benchmark

AI at Intel



Speed Up Development

Using Open AI Software

MACHINE LEARNING



TOOLKITS

App developers



Open source platform for building E2E Analytics & AI applications on Apache Spark* with distributed TensorFlow*, Keras*, BigDL



LIBRARIES

Data scientists

Python

- Scikit-learn
- Pandas
- NumPy

R

- Cart
- Random Forest
- e1071

Distributed

- MLLib (on Spark)
- Mahout



Intel-optimized Frameworks

And more framework optimizations underway including PaddlePaddle*, Chainer*, CNTK* & others



KERNELS

Library developers

Intel® Distribution for Python*

Intel distribution optimized for machine learning

Intel® Data Analytics Acceleration Library (DAAL)

High performance machine learning & data analytics library

Intel® Math Kernel Library for Deep Neural Networks (MKL-DNN)

Open source DNN functions for CPU / integrated graphics



Open source compiler for deep learning model computations optimized for multiple devices (CPU, GPU, NNP) from multiple frameworks (TF, MXNet, ONNX)

以数据为中心的世界

全球超过 **OVER**

一半 HALF 数据 **OF THE WORLD'S DATA**

创建于过去
WAS CREATED IN THE LAST
两年 **2 YEARS**

其中只有不到
的数据 **2%** **LESS THAN HAS BEEN ANALYZED**
经过了分析

大规模人工智能应用

正面临巨大的挑战



复杂性

成本

可扩展性

专有接口

数据隐私

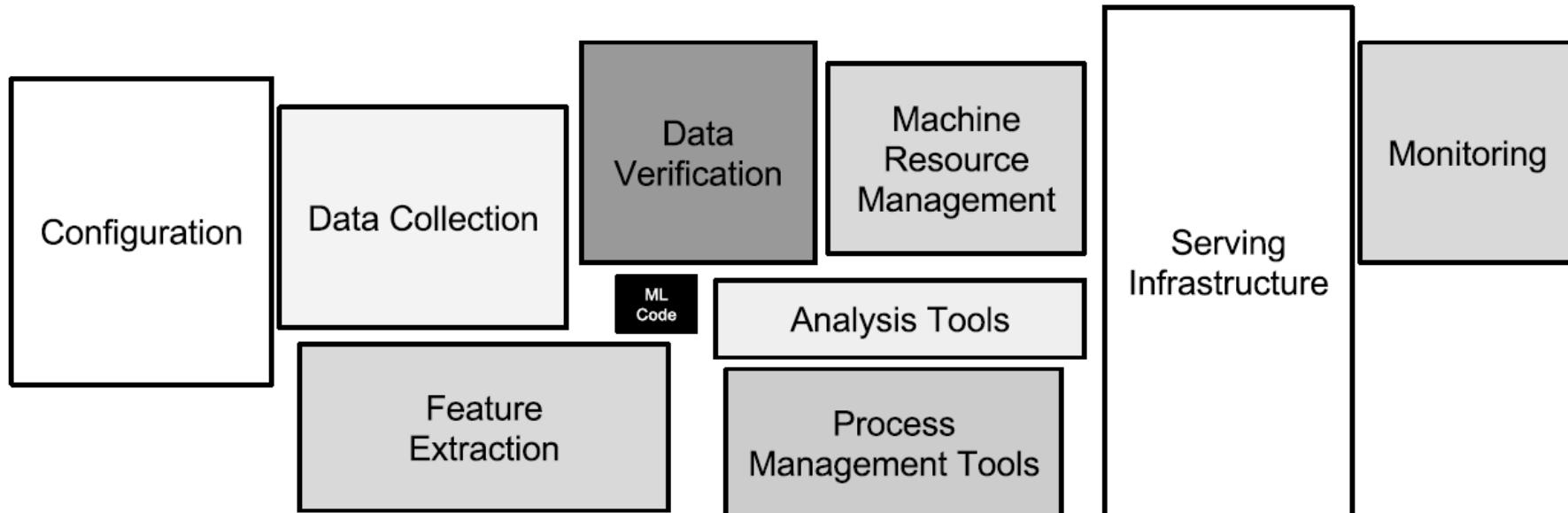


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

“Hidden Technical Debt in Machine Learning Systems”,
Sculley et al., Google, NIPS 2015

统一的数据分析及AI

获取 / 存储

清洗 / 准备

分析 / 建模

部署 / 可视化

集成的数据流水线



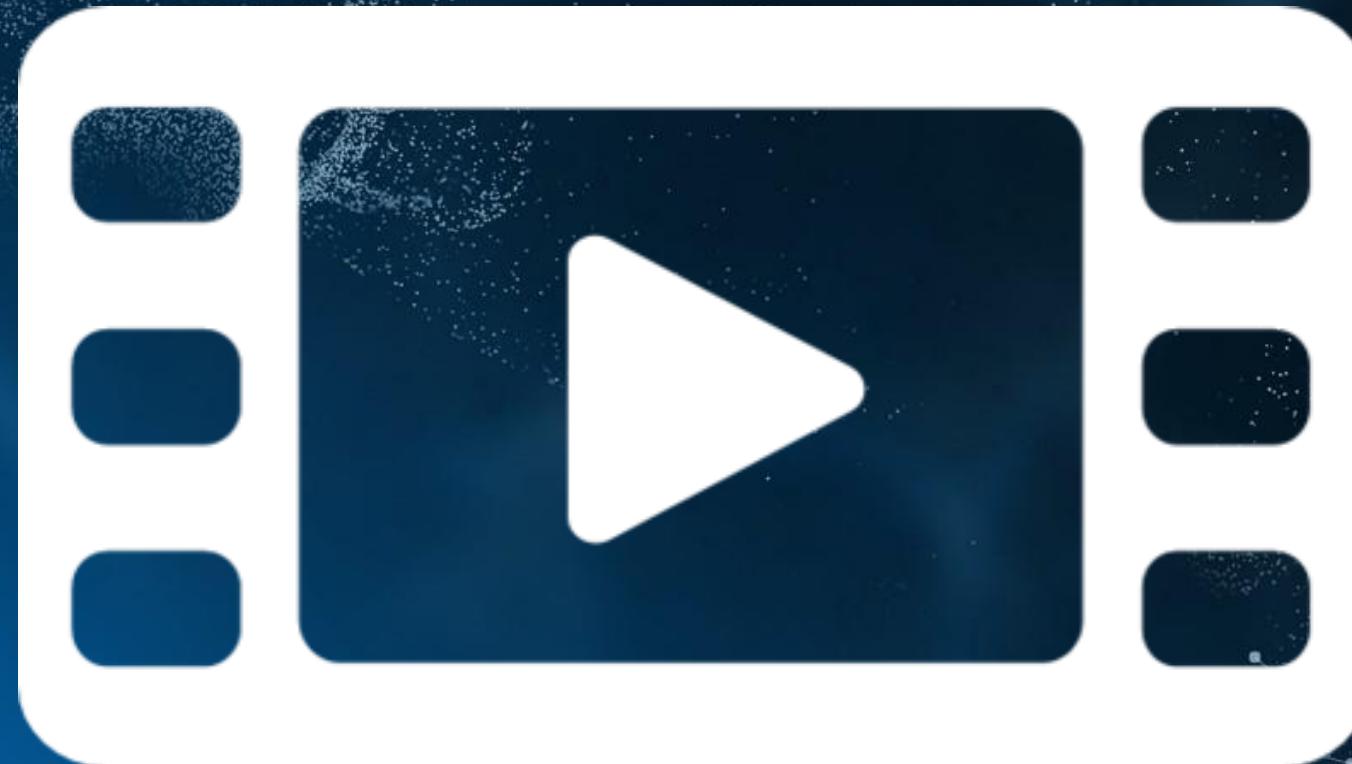
数据管理

数据分析

数据科学及人工智能

Analytics Zoo

统一的大数据分析+人工智能平台



统一的数据分析和AI流水线

端到端、从原型到生产化部署的无缝扩展

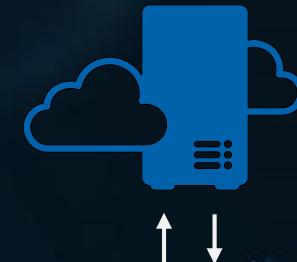
在笔记本电脑上使用
样本数据构建原型



在集群上使用历史数据
运行模型试验



在分布式生产环境中部署



生产数据流
水线

- 从笔记本电脑到分布式集群几乎**无需任何代码更改**
- **无需数据拷贝，直接访问生产大数据系统**
- 高效构建**端到端的数据分析+ AI 流水线**原型
- 无缝扩展部署到**大数据集群及生产环境**

Analytics Zoo

统一的大数据分析+人工智能平台

用户案例

Recommendation

Anomaly Detection

Text Classification

Text Matching

模型

Image Classification

Object Detection

Seq2Seq

Transformer

BERT

特征工程

image

3D image

text

time series

高级 流水线

tfpark: Distributed TensorFlow on Spark

Distributed Keras w/ autograd on Spark

nnframes: Spark Dataframes & ML
Pipelines for Deep Learning

Distributed Model Inference
(batch, streaming & online)

后端

TensorFlow*

Keras*

PyTorch*

BigDL

NLP Architect

Apache Spark*

Apache Flink*

Ray*

MKDNN

OpenVINO

Intel® Optane™ DCPMM

DL Boost (VNNI)

基于SPARK的分布式TENSORFLOW流水线

- Data wrangling and analysis using PySpark
- Deep learning model development using TensorFlow or Keras
- Distributed training / inference on Spark

```
#pyspark code
train_rdd = spark.hadoopFile(...).map(...)
dataset = TFDataSet.from_rdd(train_rdd,...)

#tensorflow code
import tensorflow as tf
slim = tf.contrib.slim
images, labels = dataset.tensors
with slim.arg_scope(lenet.lenet_arg_scope()):
    logits, end_points = lenet.lenet(images, ...)
loss = tf.reduce_mean( \
    tf.losses.sparse_softmax_cross_entropy( \
        logits=logits, labels=labels))

#distributed training on Spark
optimizer = TFOptimizer.from_loss(loss, Adam(...))
optimizer.optimize(end_trigger=MaxEpoch(5))
```

在 PySpark 程序内嵌 TensorFlow 代码

基于SPARK DATAFRAME & ML 流水线的深度学习

```
#Spark dataframe transformations
parquetfile = spark.read.parquet(...)
train_df = parquetfile.withColumn(...)

#Keras API
model = Sequential()
    .add(Convolution2D(32, 3, 3, activation='relu', input_shape=...)) \
    .add(MaxPooling2D(pool_size=(2, 2))) \
    .add(Flatten()).add(Dense(10, activation='softmax')))

#Spark ML pipeline
Estimator = NNEstimator(model, CrossEntropyCriterion()) \
    .setLearningRate(0.003).setBatchSize(40).setMaxEpoch(5) \
    .setFeaturesCol("image")
nnModel = estimator.fit(train_df)
```

在 **Spark Dataframe** 和 **ML** 流水线中，直接支持深度神经网络模型

跨行业的端到端客户案例



Alibaba Cloud
aliyun.com

Office DEPOT



Azure

CERN
openlab

阿里APACHE FLINK极客挑战赛

The screenshot shows the top navigation bar of the Alibaba Cloud website. It includes the Alibaba Cloud logo, a search bar, and a dropdown menu for '中国站'. Below the main navigation, there's a secondary navigation bar for '云栖社区' (Cloud Community) with links for '博客' (Blog), '直播' (Live Stream), '聚能聊' (Jing能 Chat), '云栖号' (Cloud Community), '专家' (Expert), '小程序云' (Cloud Mini Program), and '更多' (More). A red 'NEW' badge is visible next to the '更多' link.

云栖社区 > 博客 > 正文

首届！Apache Flink 极客挑战赛强势来袭，重磅奖项等你拿，快来组队报名啦

Ververica ① 2019-07-24 17:51:26 ② 浏览175

深度学习 大数据 性能优化 机器学习 性能 Apache 钉钉 开源大数据
流计算 大数据分析 ApacheFlink AI及大数据 实时技术

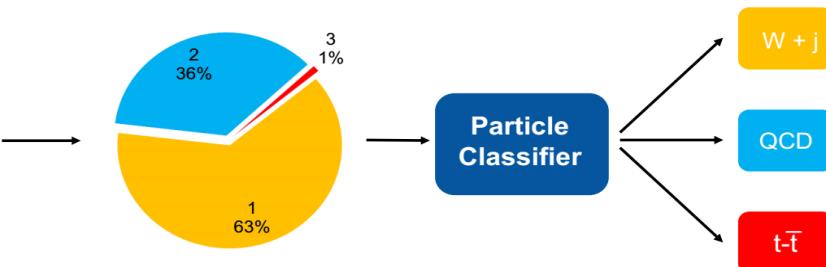
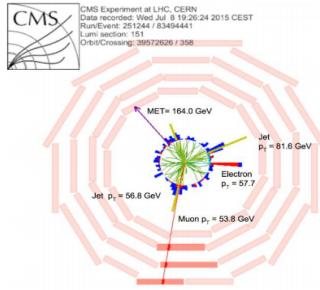
7月24日，阿里云峰会上上海开发者大会开源大数据专场，阿里巴巴集团副总裁、计算平台事业部总裁贾扬清与英特尔高级首席工程师、大数据分析和人工智能创新院院长戴金权共同发布首届Apache Flink 极客挑战赛。



这是什么（垃圾）？100个类别

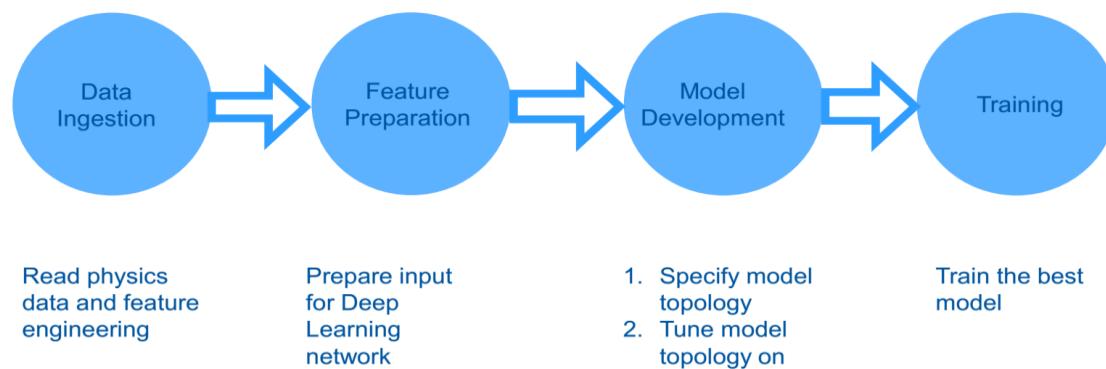


CERN*基于深度学习的高能物理粒子事件分类



高能物理数据的深度
学习流水线

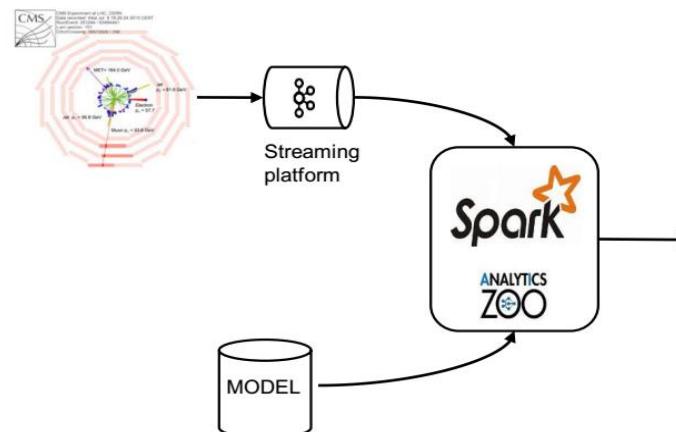
Data Pipeline



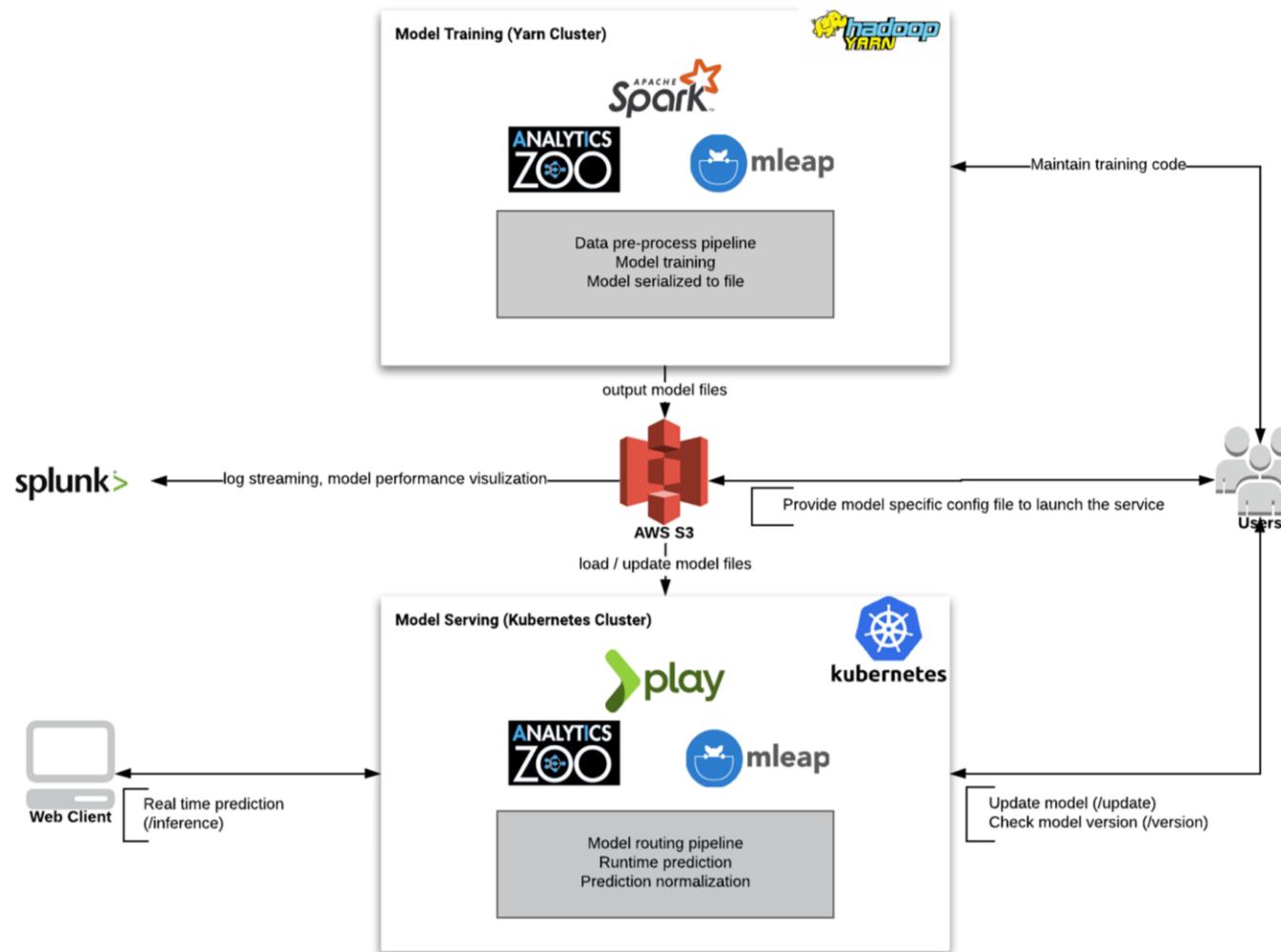
Leveraging Apache Spark and Analytics Zoo in Python Notebooks

<https://db-blog.web.cern.ch/blog/luca-canali/machine-learning-pipelines-high-energy-physics-using-apache-spark-bigdl>

<https://databricks.com/session/deep-learning-on-apache-spark-at-cerns-large-hadron-collider-with-intel-technologies>

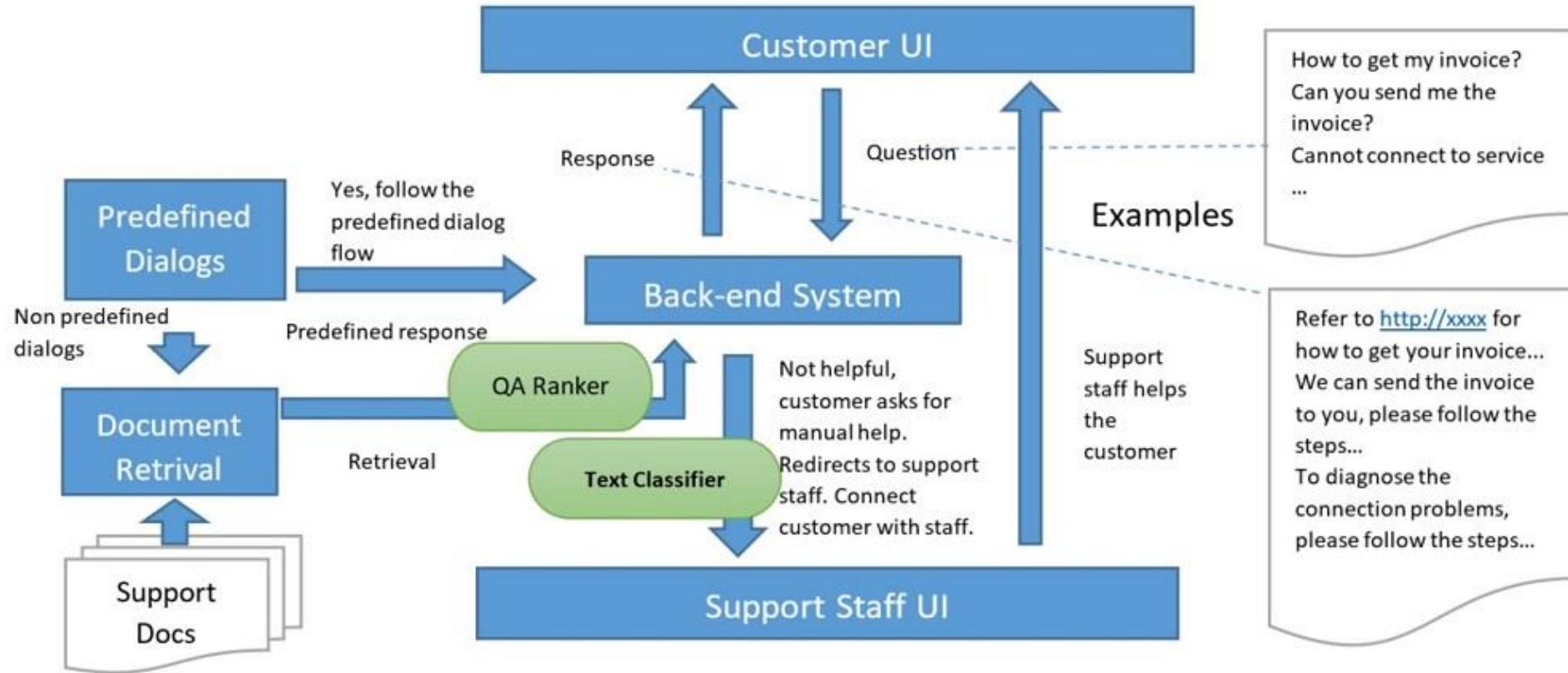


OFFICE DEPOT*: 基于用户 SESSION 行为的产品推荐



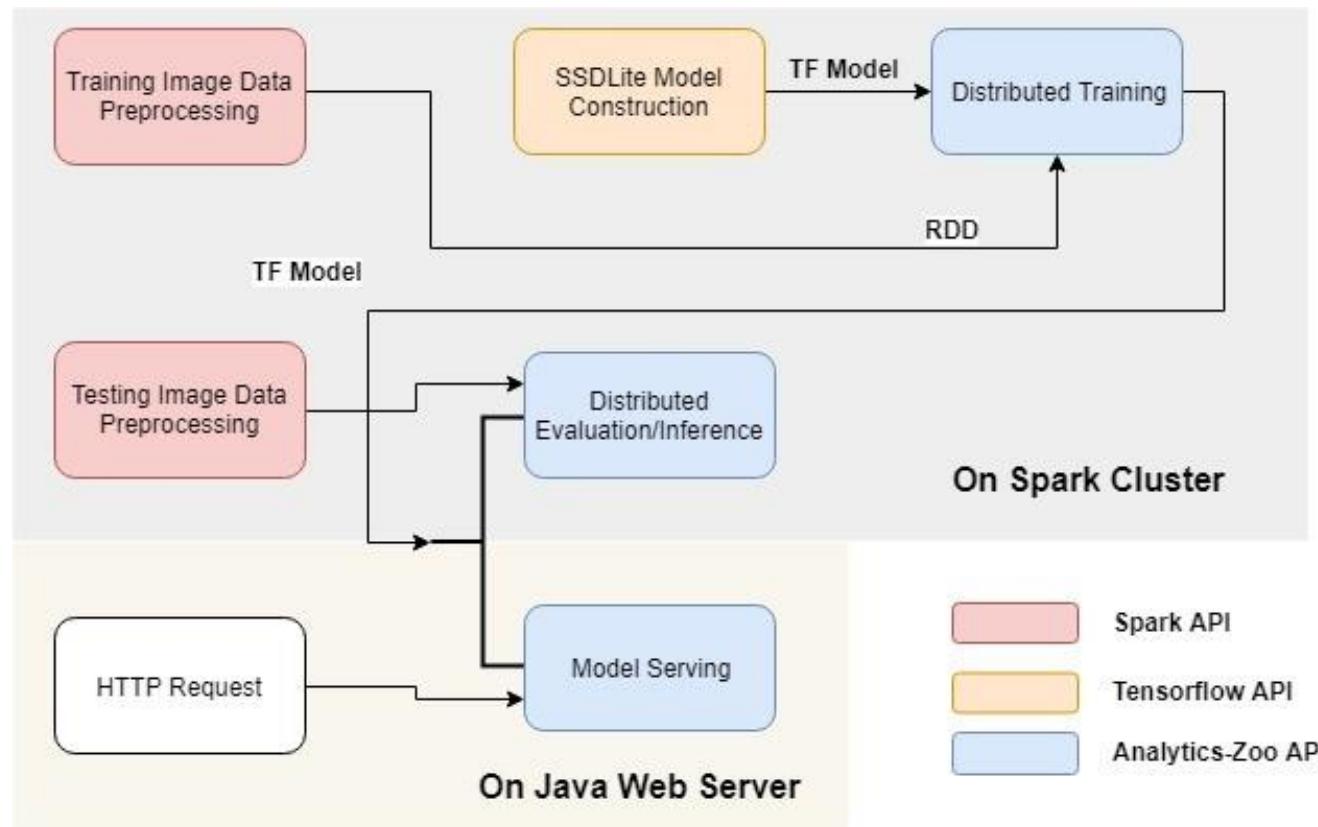
<https://software.intel.com/en-us/articles/real-time-product-recommendations-for-office-depot-using-apache-spark-and-analytics-zoo-on>
<https://conferences.oreilly.com/strata/strata-ca-2019/public/schedule/detail/73079>

微软AZURE CHATBOT



<https://software.intel.com/en-us/articles/use-analytics-zoo-to-inject-ai-into-customer-service-platforms-on-microsoft-azure-part-1>
<https://www.infoq.com/articles/analytics-zoo-qa-module/>

美的*: 工业视觉检测云平台

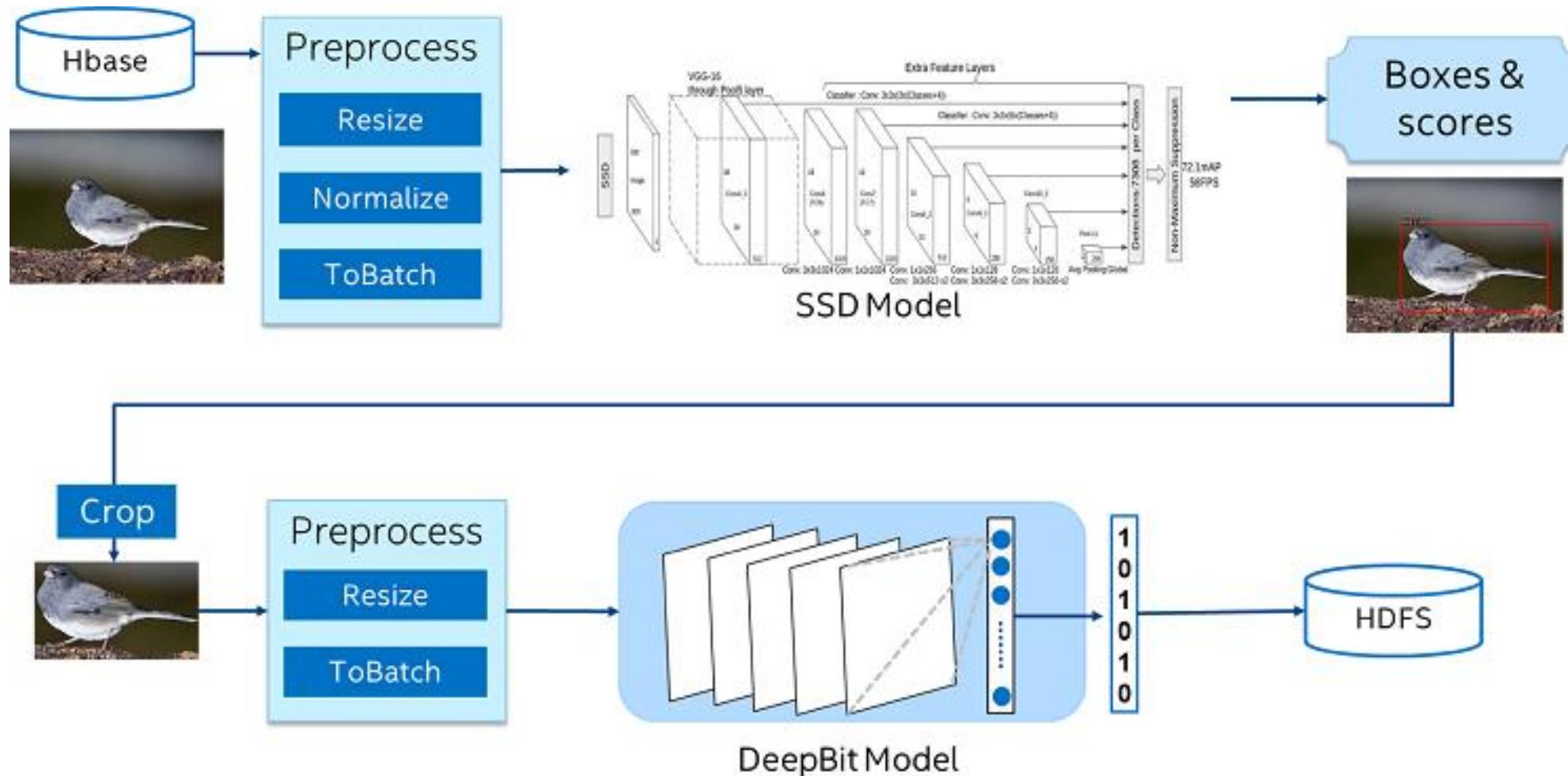


<https://software.intel.com/en-us/articles/industrial-inspection-platform-in-midea-and-kuka-using-distributed-tensorflow-on-analytics>
<https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/midea-case-study.html>

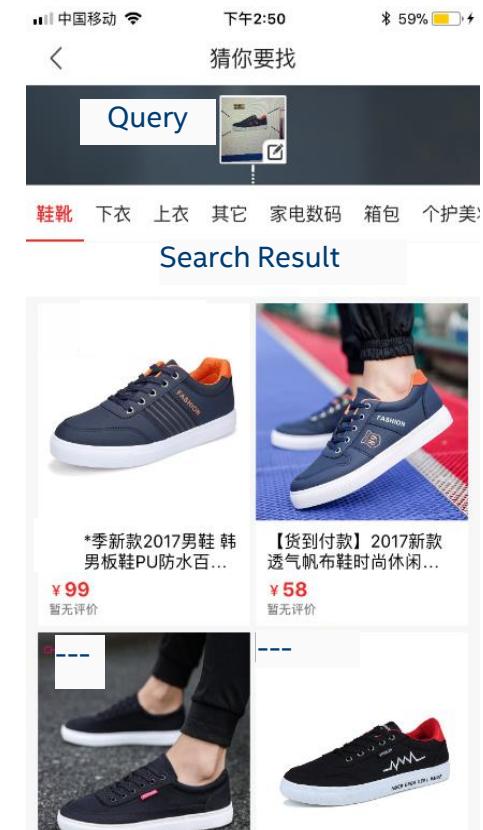
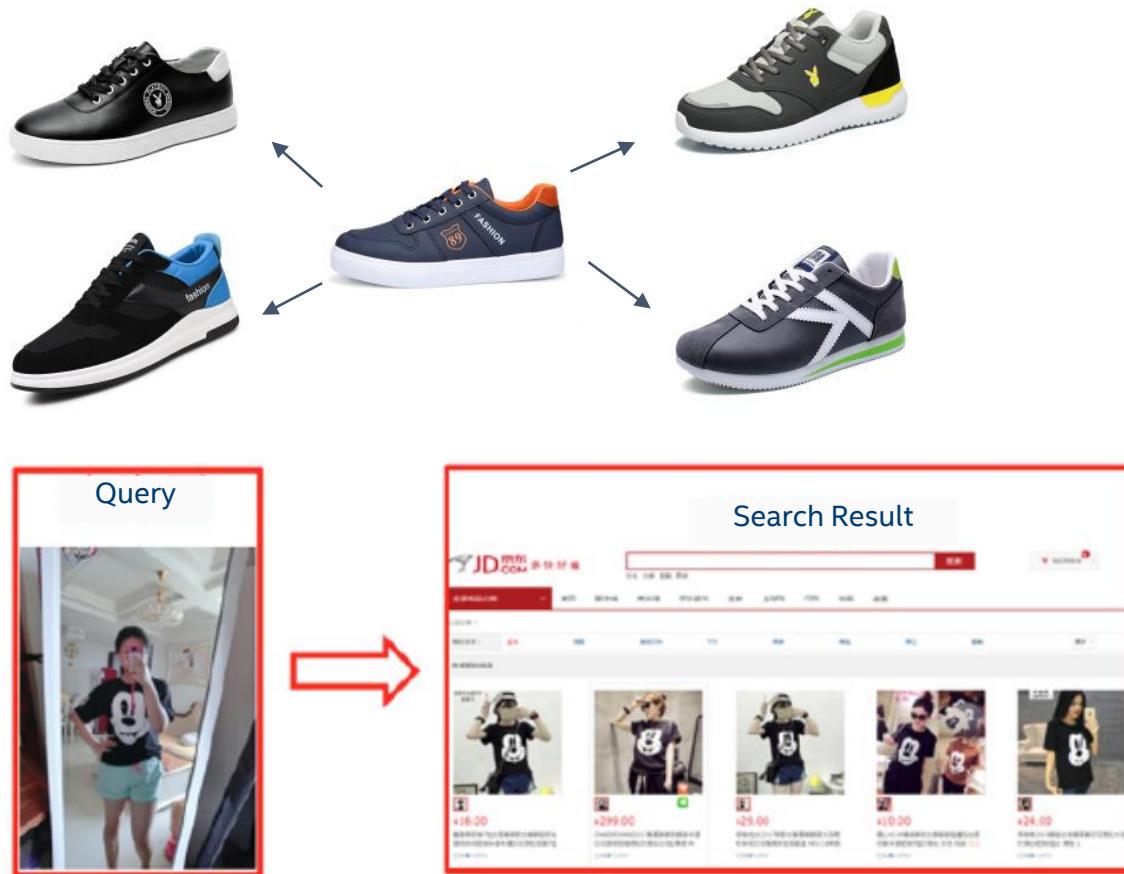
美的*: 工业视觉检测云平台



JD.COM：目标检测和图像特征提取

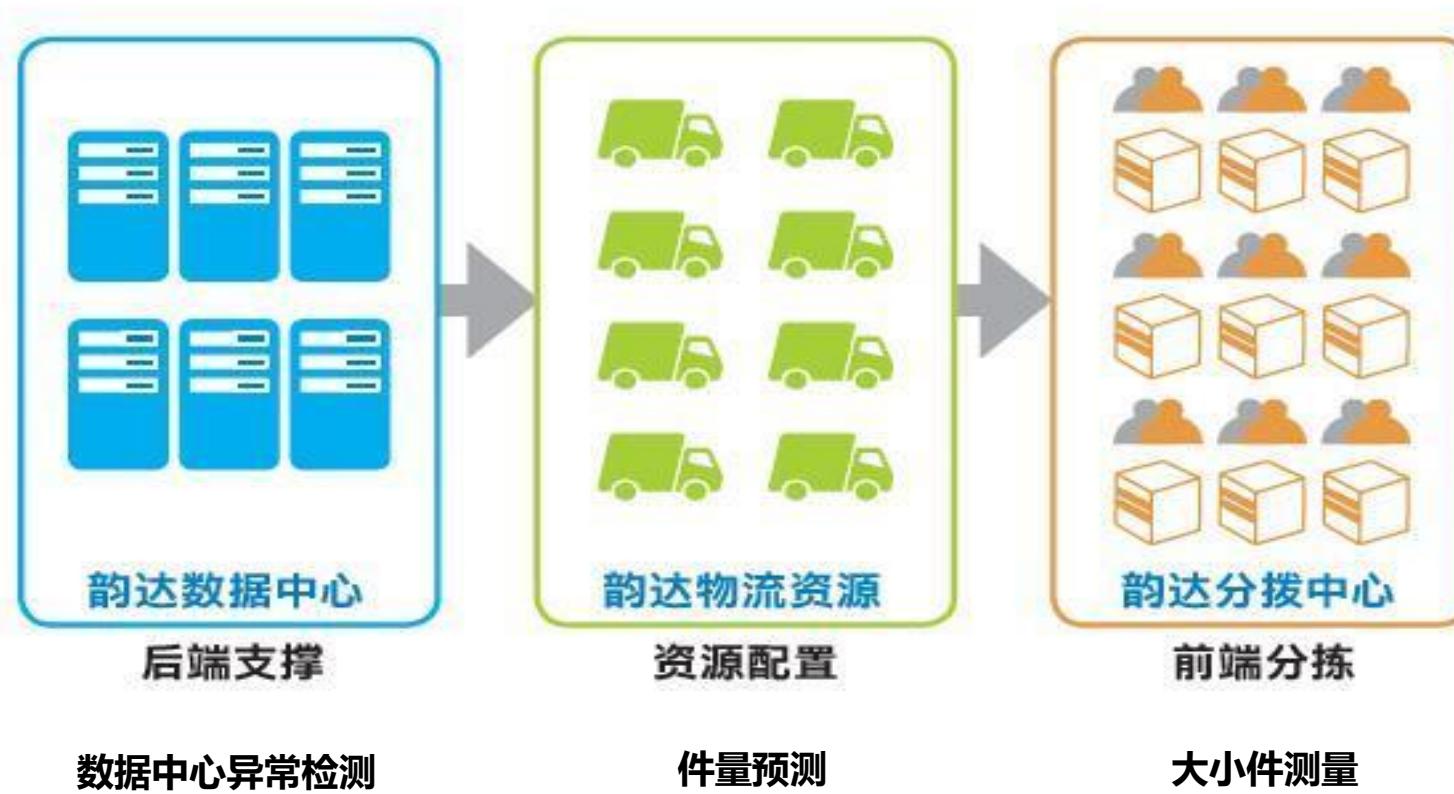


相似图片搜索



Source: "Bringing deep learning into big data analytics using BigDL", Xianyan Jia and Zhenhua Wang, Strata Data Conference Singapore 2017

韵达*: 基于AI提升快递物流系统运转效率



<https://www.intel.cn/content/www/cn/zh/analytics/artificial-intelligence/yunda-brings-quality-change-to-the-express-delivery-industry.html>

阿里云E-MAPREDUCE SERVICE

在阿里云EMR上使用ANALYTICS ZOO 集群的创建

The screenshot shows the 'Create Cluster' wizard in the Alibaba Cloud E-MapReduce console. The current step is 'Software Configuration'. The interface includes tabs for 'Overview', 'Cluster Management' (selected), 'Data Development' (New), 'Metadata Management', 'Monitoring Dashboard' (Beta), 'System Maintenance', 'Operation Log', 'Help', and 'Old Job Scheduling'. The 'Cluster Management' tab has sub-sections: 'Create Cluster' (selected), 'Edit Cluster', 'Delete Cluster', and 'Cluster Status'. The 'Create Cluster' section has four steps: 'Software Configuration' (highlighted in blue), 'Hardware Configuration', 'Basic Configuration', and 'Confirmation'. The 'Software Configuration' step is titled 'Version Configuration' and contains the following fields:

- Product Version: EMR-3.15.0-yh_pre
- Cluster Type: Data Science (radio button selected, others: Hadoop, Druid, Kafka, ZooKeeper)
- Mandatory Services:
 - ANALYTICS-ZOO (0.2.0) (selected)
 - Jupyter (4.4.0)
 - Knox (0.13.0-0.0.3)
 - ApacheDS (2.0.0)
 - Tensorflow On YARN (1.0.0)
 - Tensorflow (1.8.0)
 - Zeppelin (0.8.0)
 - Spark (2.3.1-1.1.2)
 - YARN (2.7.2-1.3.1)
 - HDFS (2.7.2-1.3.1)
 - ZooKeeper (3.4.13)
 - Ganglia (3.7.2)
- Optional Services:
 - HUE (4.1.0) (selected)
 - Hive (2.3.3-1.0.2)
- High Security Mode: A toggle switch is off.
- Custom Software Configuration: A toggle switch is off.

<https://partners-intl.aliyun.com/help/doc-detail/93155.htm>

And Many More

Not a full list

TECHNOLOGY



CLOUD SERVICE PROVIDERS



END USERS



software.intel.com/AlonBigData

Analytics Zoo 近期将支持的新功能



- 对 Flink 的更多支持
- Cluster Serving 分布式推理 (Part 2 of this session)

PyTorch



RAY on

- 基于 Apache Spark 的分布式

- 直接在 Hadoop 集群上支持

- <https://medium.com/riselab/rayonspark-running-emerging-ai-applications-on-big-data-clusters-with-ray-and-analytics-zoo-923e0136ed6a>

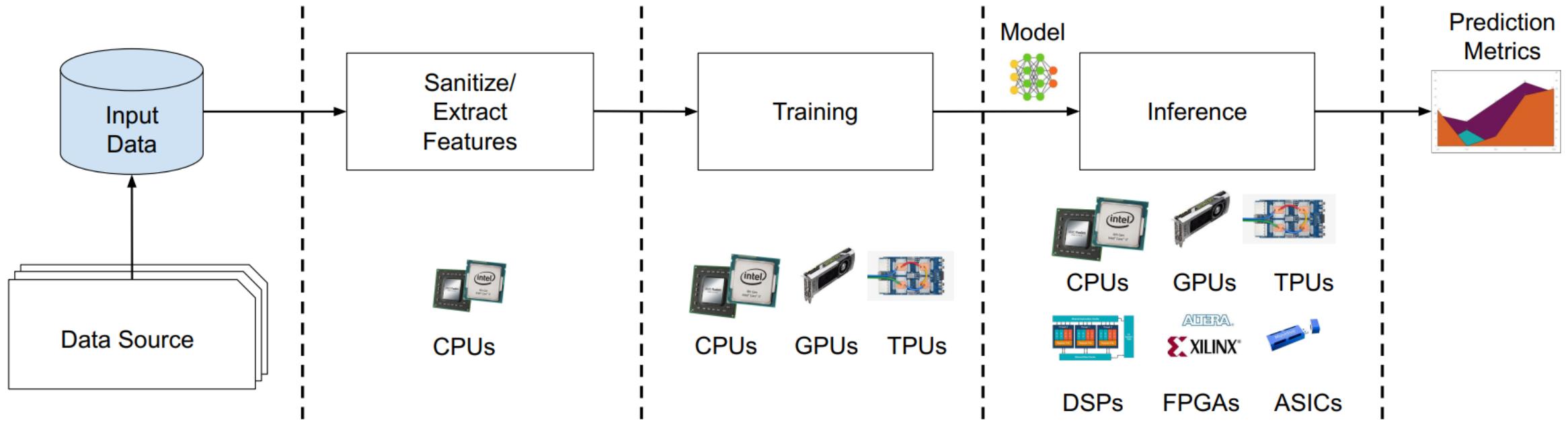
- AutoML 支持

- 对时间序列数据 (如异常检测或趋势预测) 进行自动的特征生成、模型选择和超参调优

The background image is a wide-angle, aerial photograph of a modern urban landscape at night. A large cable-stayed bridge spans a wide river, its towers illuminated against a dark sky. In the foreground, a complex multi-level highway interchange is visible, with streaks of light from moving vehicles creating dynamic, colorful lines. The city skyline in the distance is dotted with numerous lit buildings and skyscrapers. The overall atmosphere is one of a bustling, technologically advanced metropolis.

CLUSTER SERVING WITH
ANALYTICS ZOO

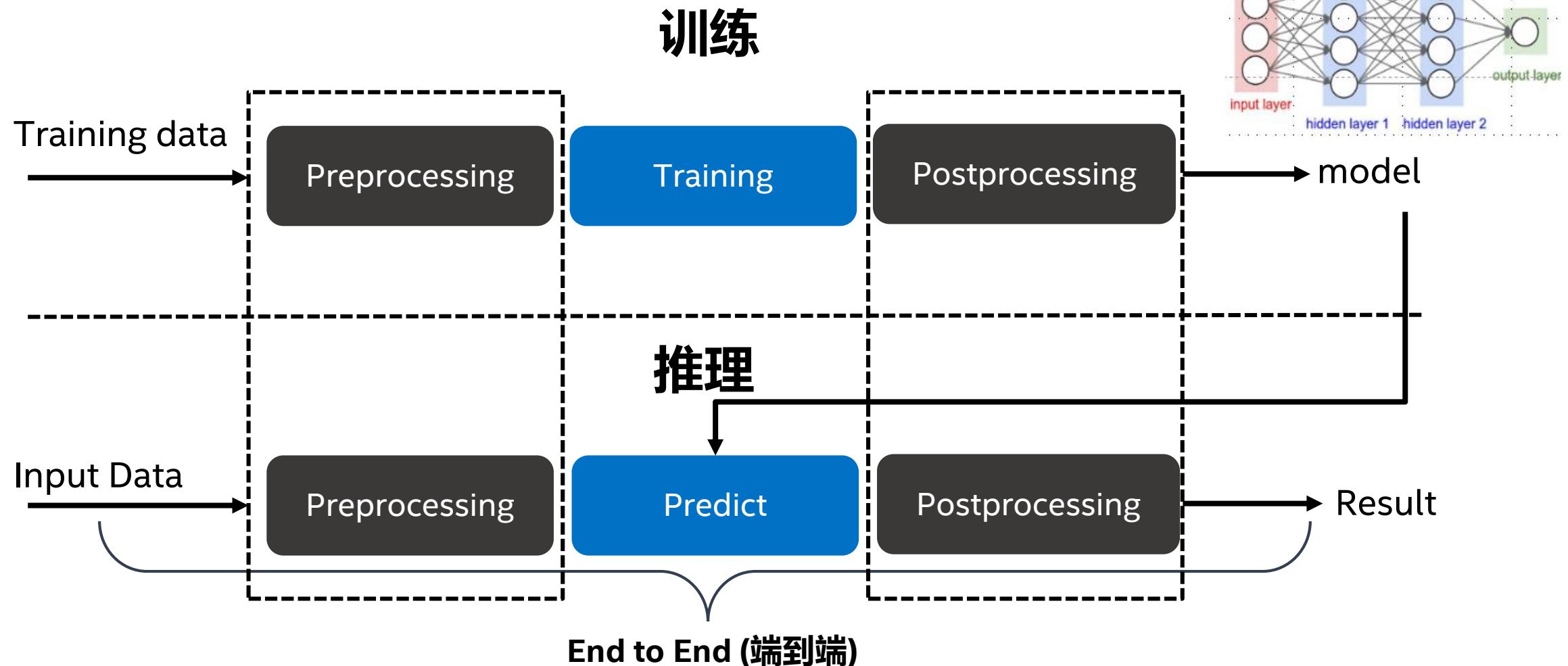
MACHINE/DEEP LEARNING PIPELINE



<https://mlperf.org/>

VJ Reddi, 2019, MLPerf Inference Benchmark

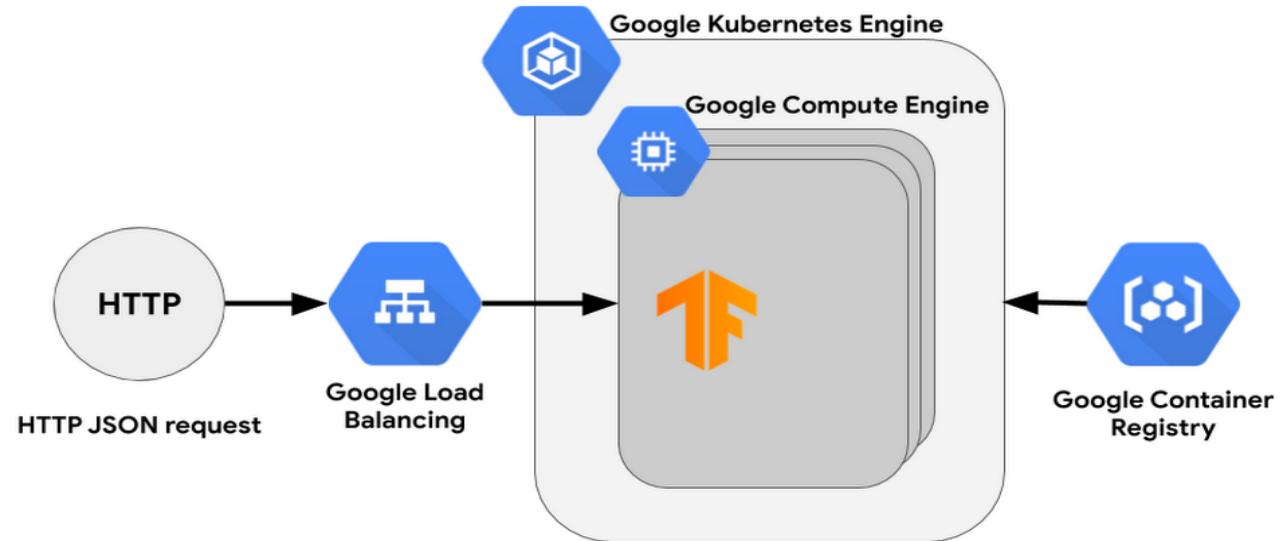
MACHINE/DEEP LEARNING PIPELINE



MODEL SERVING

DL Framework provided Serving

- TensorFlow Serving
- PyTorch Serving
- MXNet Serving
- PaddlePaddle Serving
- ...



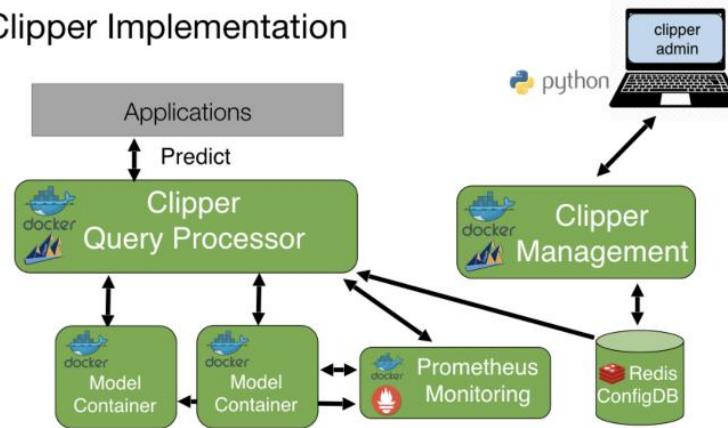
<https://cloud.google.com/blog/products/ai-machine-learning/how-to-serve-deep-learning-models-using-tensorflow-2-0-with-cloud-functions>

MODEL SERVING

General Serving framework

- OpenVINO Serving
- TensorRT Inference Server
- Kubeflow
- MLFlow
- Clipper
- Seldon
-

Clipper Implementation



<http://clipper.ai/>

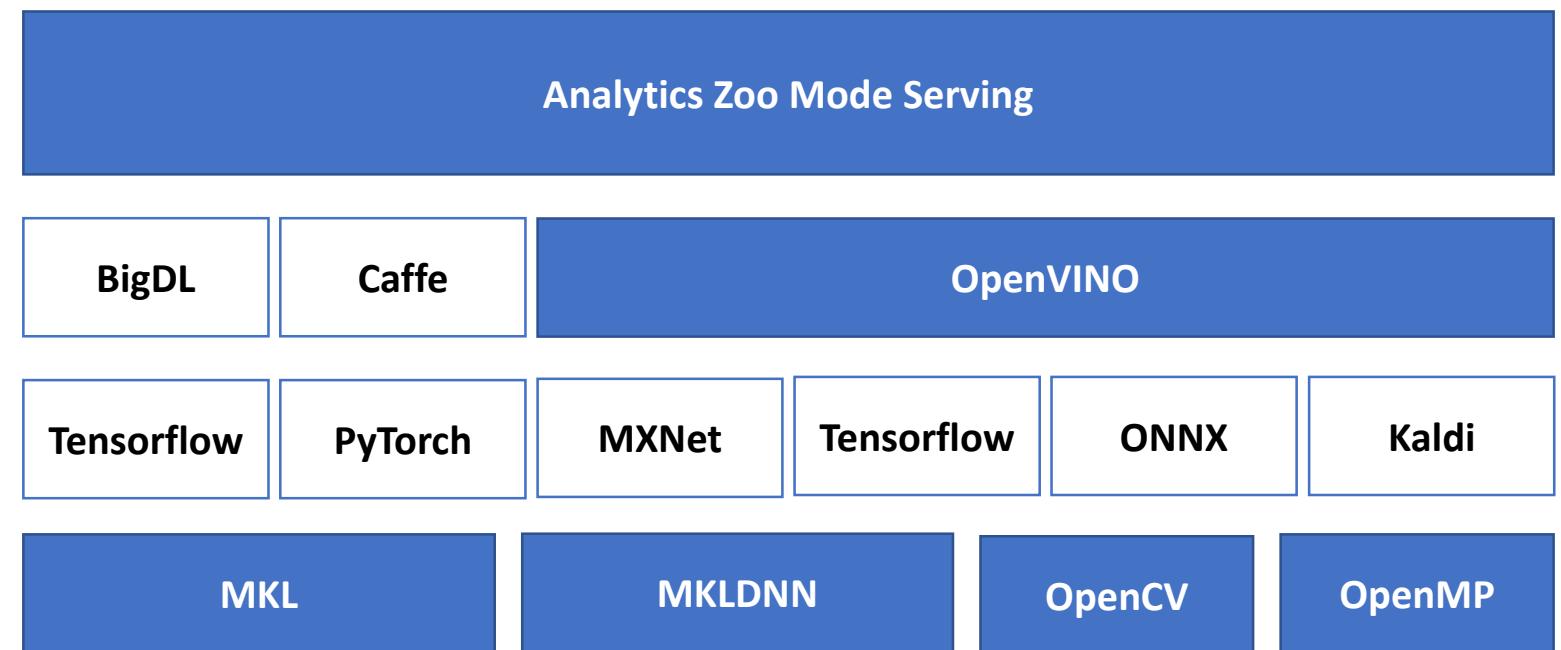


<https://github.com/kubeflow/kfserving>

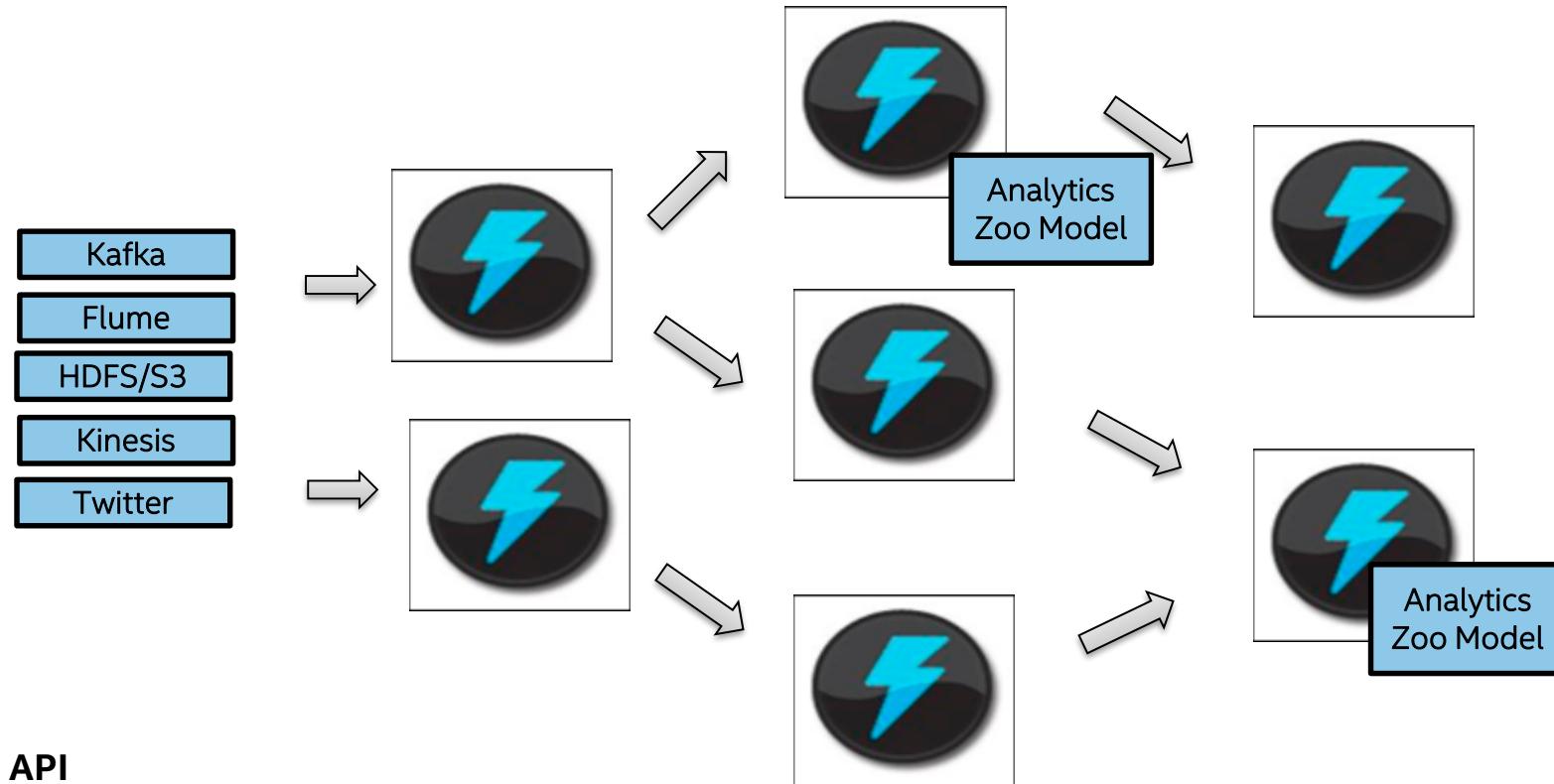
ANALYTICS-ZOO MODEL SERVING

支持多种深度学习框架的模型

- BigDL
- Caffe
- Tensorflow
- PyTorch
- OpenVINO
- Kaldi

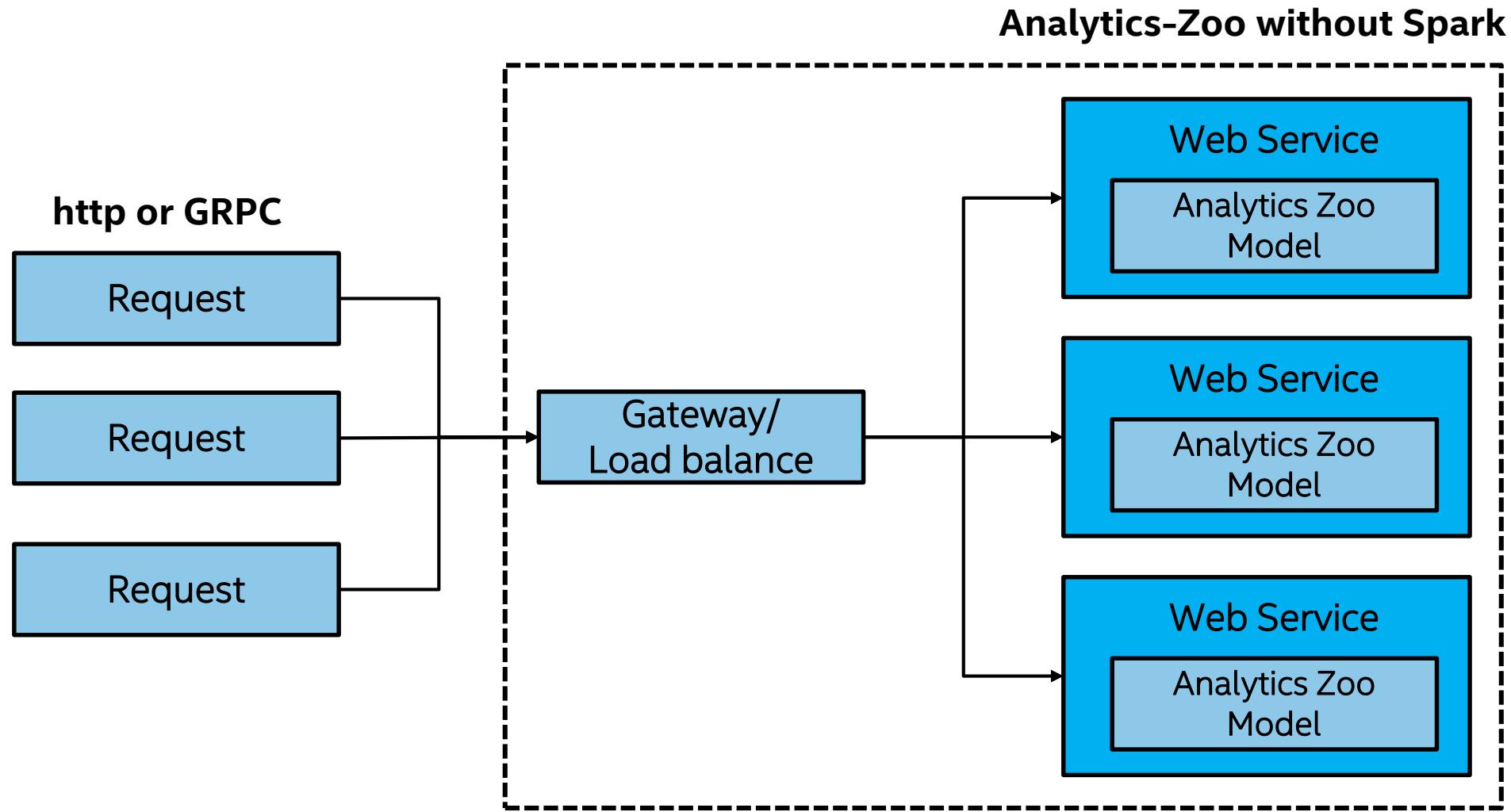


分布式、实时(流式)模型推理流水线



- 纯 Java 或 Python API
- 支持 Flink、Kafka、Storm、Web Service 等
- OpenVINO 和 DL Boost (VNNI) 加速

基于WEB SERVICE的分布式MODEL SERVING



基于WEB SERVICE的分布式MODEL SERVING API

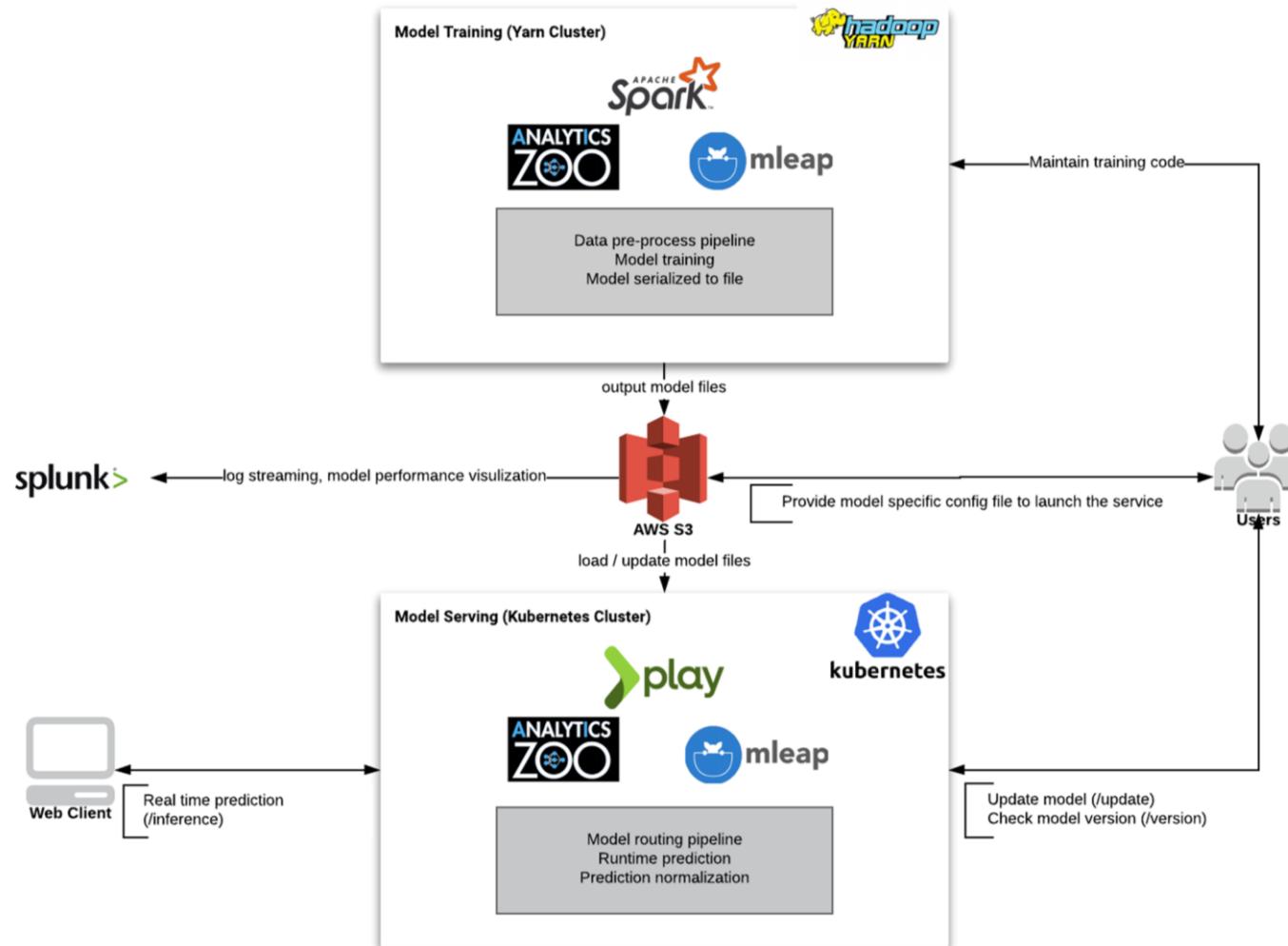
```
import com.intel.analytics.zoo.pipeline.inference.AbstractInferenceModel;

public class TextClassification extends AbstractInferenceModel {
    public RankerInferenceModel(int concurrentNum) {
        super(concurrentNum);
    }
    ...
}

public class ServingExample {
    public static void main(String[] args) throws IOException {
        TextClassification model = new TextClassification();
        model.load(modelPath, weightPath);

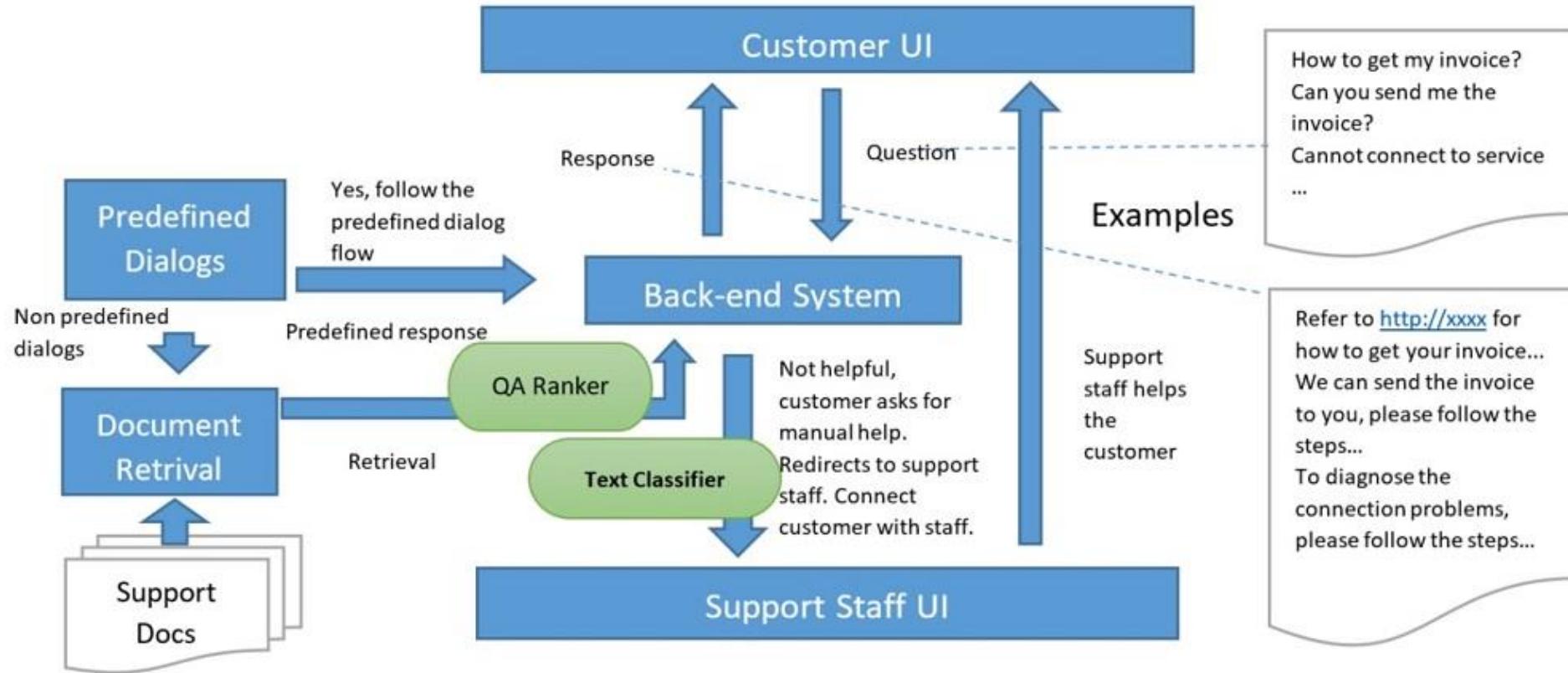
        texts = ...
        List<JTensor> inputs = preprocess(texts);
        for (JTensor input : inputs) {
            List<Float> result = model.predict(input.getData(), input.getShape());
            ...
        }
    }
}
```

OFFICE DEPOT*: 基于用户 SESSION 行为的产品推荐



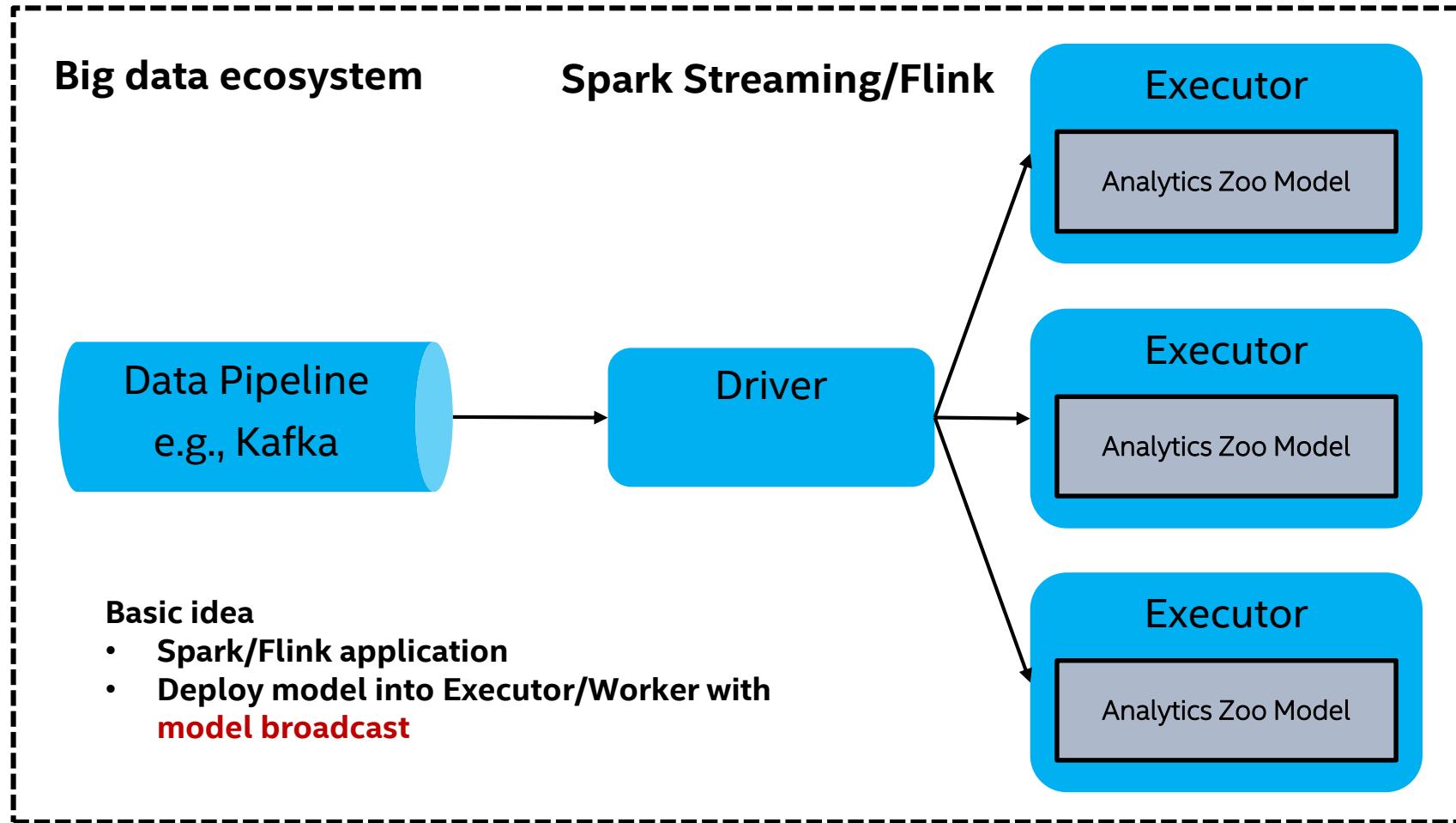
<https://software.intel.com/en-us/articles/real-time-product-recommendations-for-office-depot-using-apache-spark-and-analytics-zoo-on>
<https://conferences.oreilly.com/strata/strata-ca-2019/public/schedule/detail/73079>

微软AZURE CHATBOT

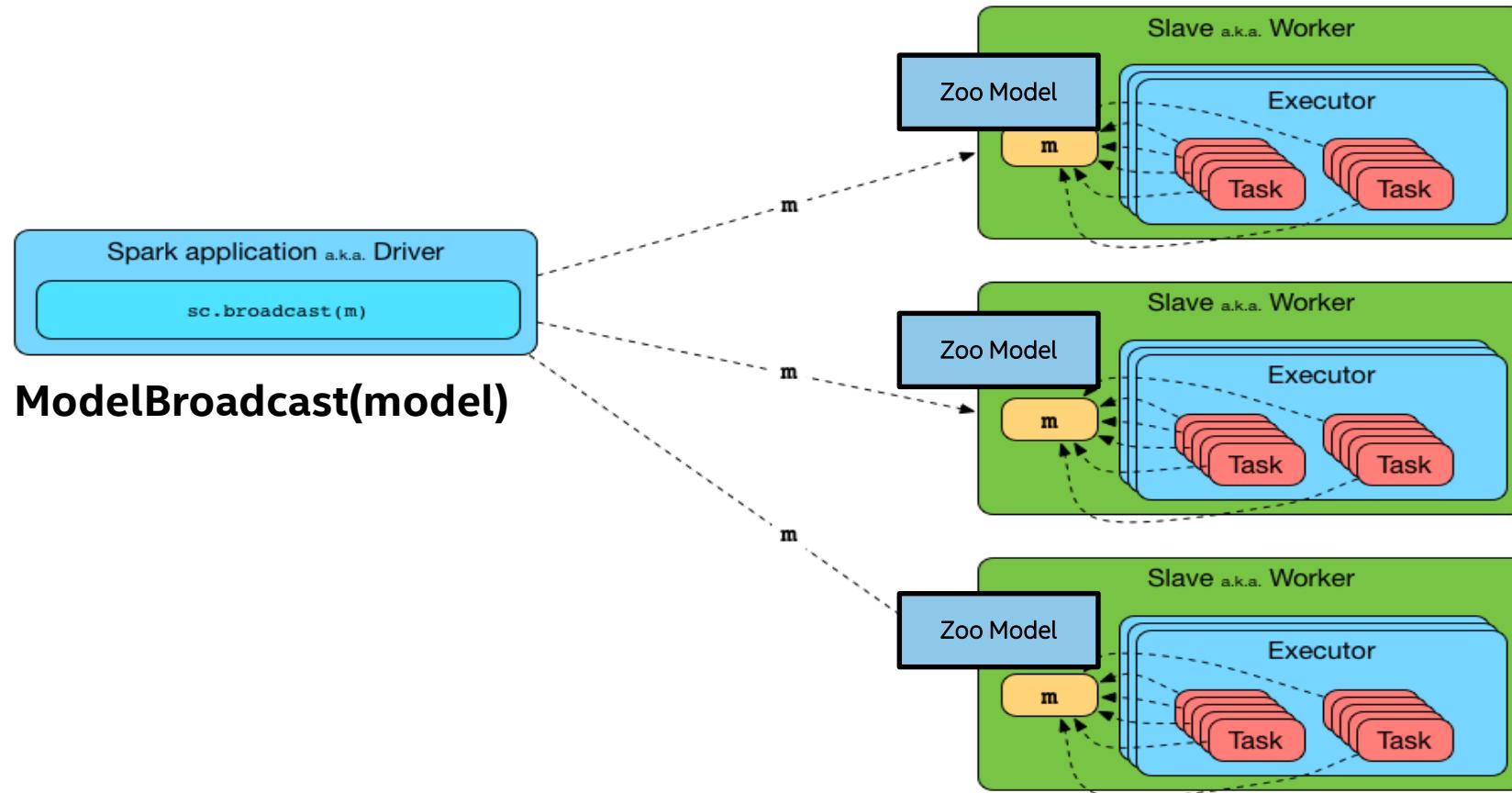


<https://software.intel.com/en-us/articles/use-analytics-zoo-to-inject-ai-into-customer-service-platforms-on-microsoft-azure-part-1>
<https://www.infoq.com/articles/analytics-zoo-qa-module/>

基于STREAMING的MODEL SERVING



基于STREAMING的MODEL SERVING(SPARK)



基于STREAMING的MODEL SERVING

```
sc = init_nncontext("Streaming Object Detection Example")
ssc = StreamingContext(sc, 3)
# Spark Streaming
lines = ssc.textFileStream(args.streaming_path)

# Analytics-Zoo ObjectDetector API
model = ObjectDetector.load_model(args.model)

def predict(batch_path):
    if batch_path.getNumPartitions() == 0:
        return
    image_set = DistributedImageSet(batch_path.map(read_image_file))
    output = model.predict_image_set(image_set)
    # Save to output
    config = model.get_config()
    visualizer = Visualizer(config.label_map(), encoding="jpg")
    visualizer(output).get_image(to_chw=False) \
        .foreach(lambda x: write_image_file(x, args.output_path))

lines.foreachRDD(predict)
# Start the computation
ssc.start()
# Wait for the computation to terminate
ssc.awaitTermination()
```

阿里APACHE FLINK极客挑战赛

The screenshot shows the top navigation bar of the Alibaba Cloud website. It includes the Alibaba Cloud logo, a search bar, and a dropdown menu for '中国站'. Below the main navigation, there's a secondary navigation bar for '云栖社区' (Cloud Community) with links for '博客' (Blog), '直播' (Live Stream), '聚能聊' (Jing能 Chat), '云栖号' (Cloud Community), '专家' (Expert), '小程序云' (Cloud Mini Program), and '更多' (More). A red 'NEW' badge is visible next to the '更多' link.

云栖社区 > 博客 > 正文

首届！Apache Flink 极客挑战赛强势来袭，重磅奖项等你拿，快来组队报名啦

Ververica ① 2019-07-24 17:51:26 ② 浏览175

深度学习 大数据 性能优化 机器学习 性能 Apache 钉钉 开源大数据
流计算 大数据分析 ApacheFlink AI及大数据 实时技术

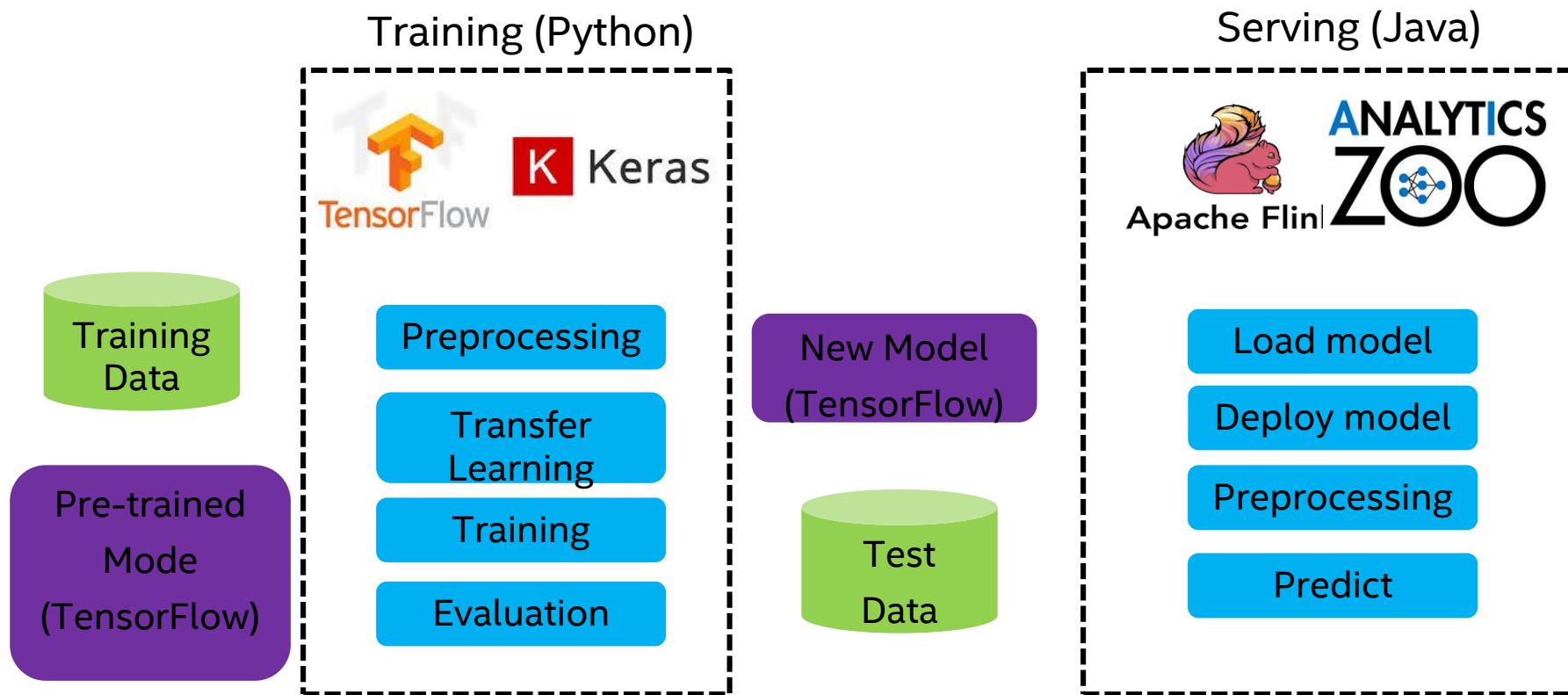
7月24日，阿里云峰会上上海开发者大会开源大数据专场，阿里巴巴集团副总裁、计算平台事业部总裁贾扬清与英特尔高级首席工程师、大数据分析和人工智能创新院院长戴金权共同发布首届Apache Flink 极客挑战赛。



这是什么（垃圾）？100个类别

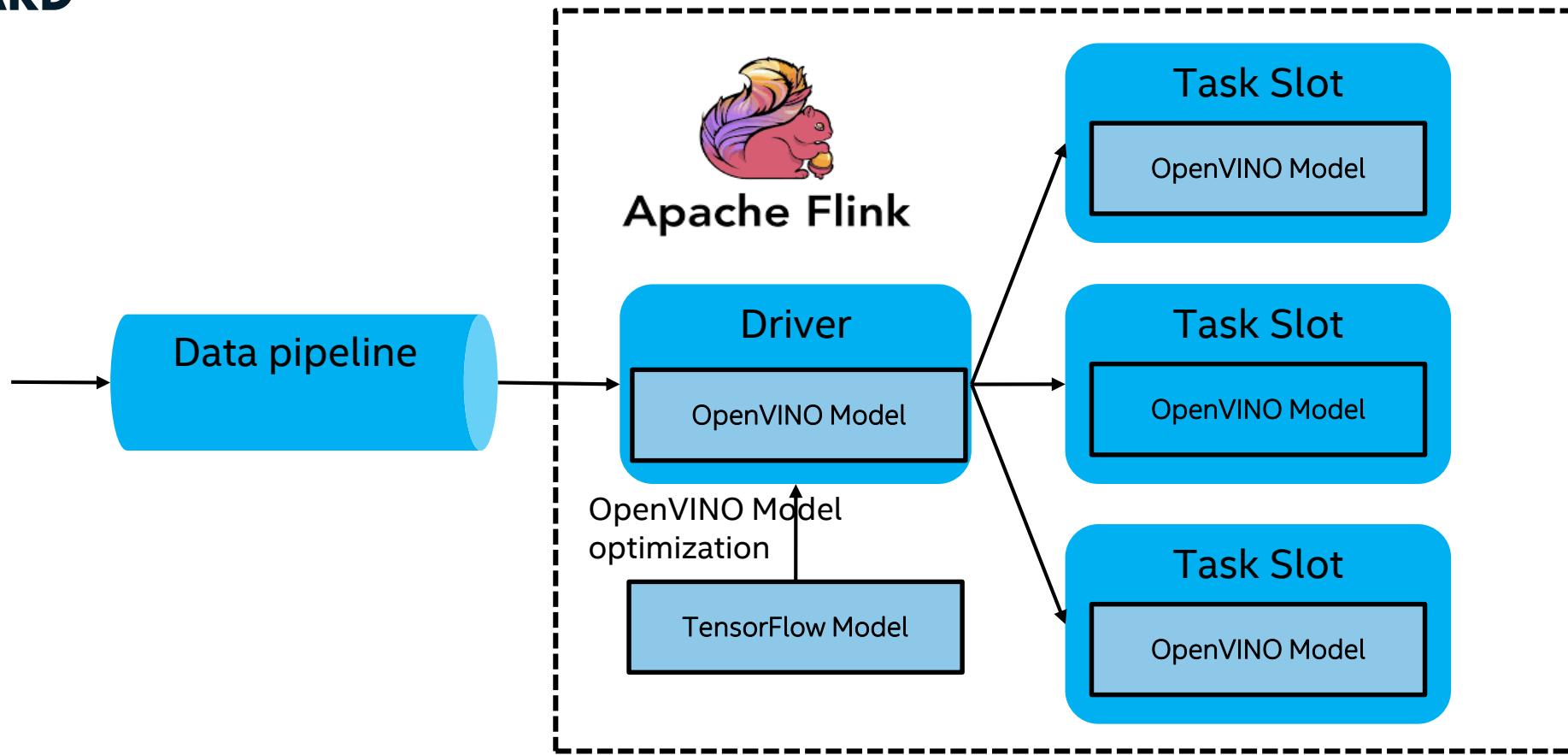


阿里APACHE FLINK极客挑战赛



基于STREAMING的MODEL SERVING (FLINK)

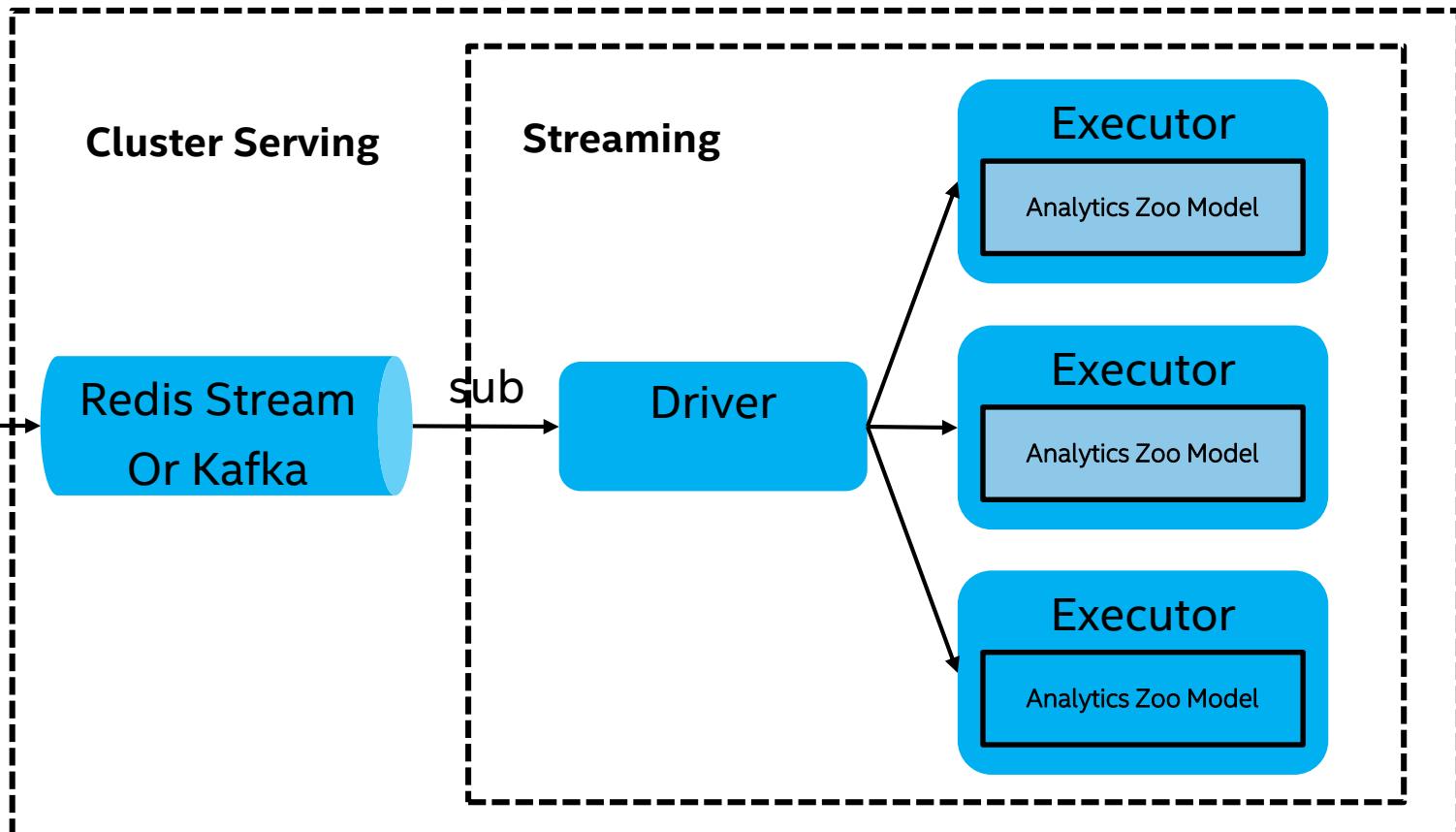
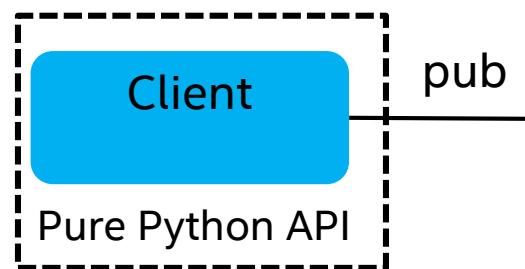
FLINK FORWARD 在Flink*上使用Analytics Zoo进行大数据分析与深度学习模型推理的架构与实践



ANALYTICS-ZOO CLUSTER SERVING

Pub/Sub with Message Queue

- *Real-time*
- *Simplified Python API*
- *Easy to deploy (config.yaml & one line command)*



ANALYTICS-ZOO CLUSTER SERVING

部署

- ✓ 一个本地节点或者一个Docker容器
- ✓ 已有的 Yarn/Spark/Flink (or K8s) 集群

使用



- 1 一条命令:
- 启动Docker容器以及Zoo Cluster Serving

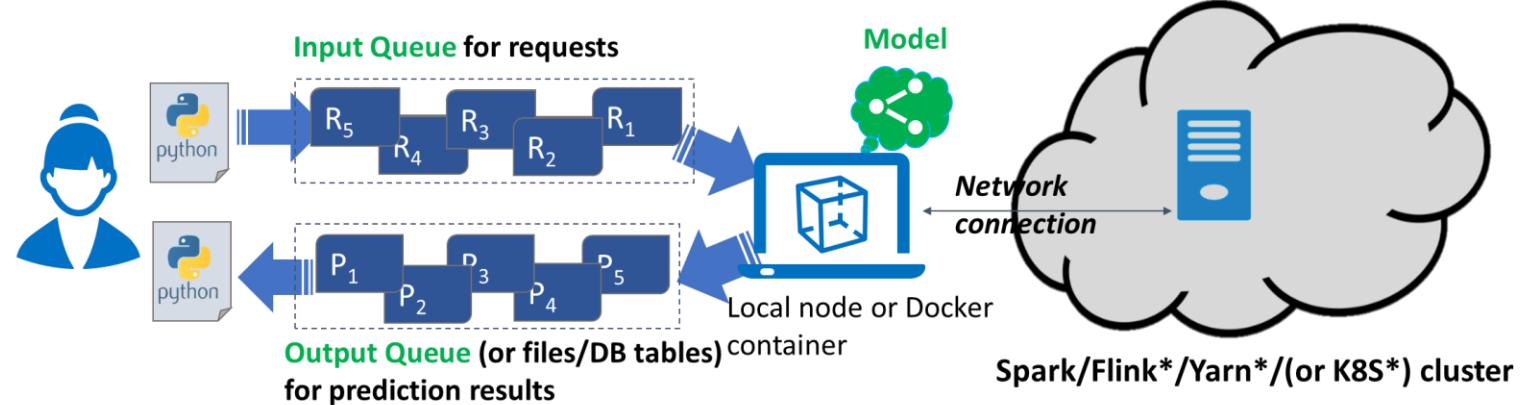
- 此命令指定:
 - 输入 和 输出 的队列名字
 - 模型 的文件路径
 - 预/后处理 的文件路径
 - 集群 的访问路径



- 2 一个简单的Python脚本:
- 将请求数据发送到 Input Queue
 - 从 Output Queue (或文件/数据库)获得推理结果

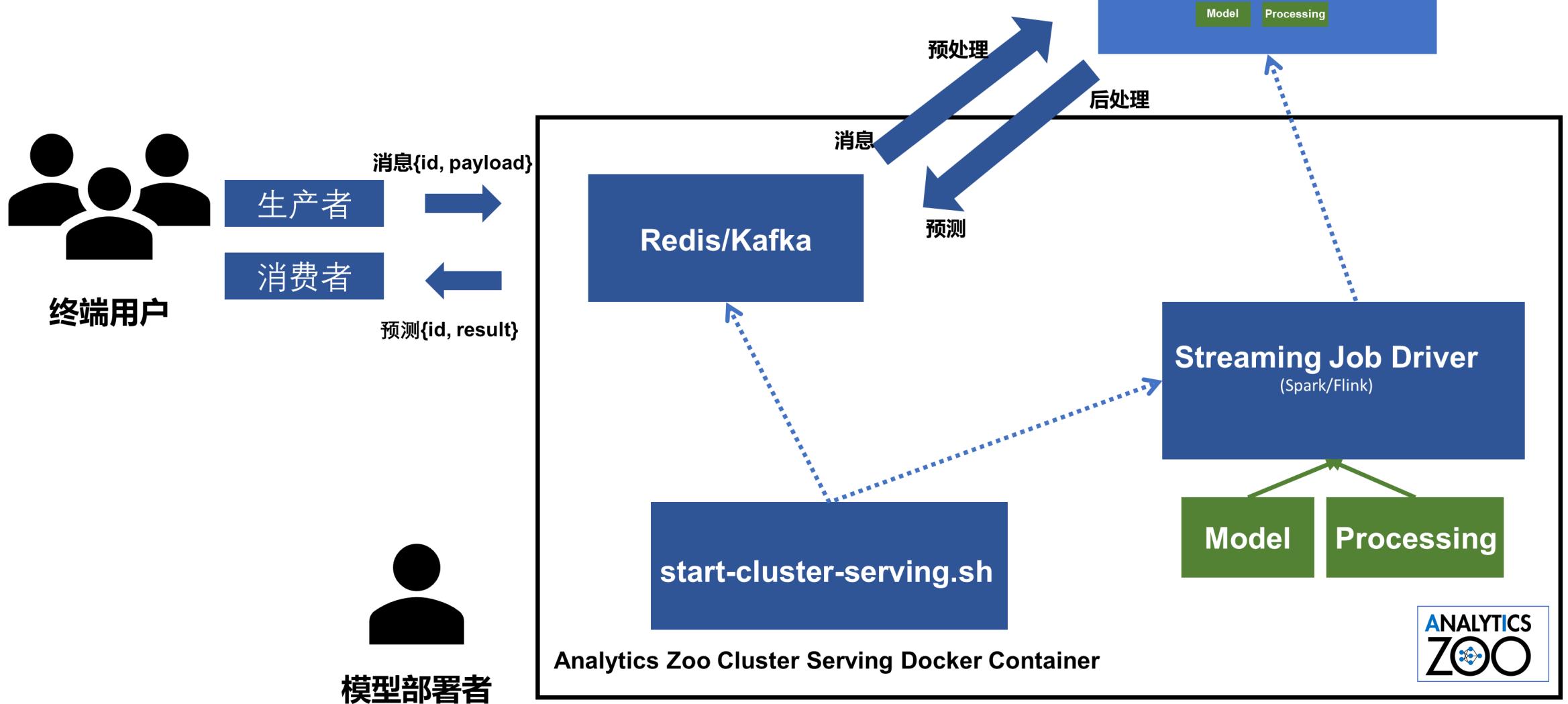


- 3 Analytics Zoo 在集群上自动执行分布式、实时 (流式) 模型推理
- 支持 TensorFlow*, Keras*, PyTorch*, Caffe*, BigDL 和 OpenVINO 的模型, 可使用 Int8 加速
 - 通过Spark/Flink* 线性扩展



- ✓ 可扩展的分布式推理由Analytics Zoo托管
- ✓ 用户无需为开发和部署复杂的分布式推理方案而费心

ANALYTICS-ZOO CLUSTER SERVING



ANALYTICS-ZOO CLUSTER SERVING

Pure Python Client

```
from zoo.serving.client.utils.helpers import RedisQueue
import cv2

if __name__ == "__main__":
    # if you do not specify config_path, it will use default config
    config_path = "/path/to/analytics-zoo-cluster-serving/config.yaml"
    redis_queue = RedisQueue(config_path)

    # following lines demonstrate how to push data to redis
    img_path = "/path/to/image"
    # pre-processing in client is optional
    img = cv2.imread(img_path)
    img = cv2.resize(img, (224, 224))
    redis_queue.enqueue_image(img)

    # following lines demonstrate how to get data from redis
    redis_queue.get_results("result:*)
```

ANALYTICS-ZOO CLUSTER SERVING

YAML Configuration

```
## Analytics-zoo Cluster Serving
model:
  # model path must be set
  path: resources
data:
  # redis address
  src: XXXXXX:6379
  shape: 3, 224, 224

  # default, image-classification
  task:

spark:
  # default, local[*]
  master: spark://XXXX:7077

params:
  # default, 4
  batch_size: 8

  # default, mkldnn
  engine_type
```

ANALYTICS-ZOO CLUSTER SERVING

User Customized Serving Pipeline

```
from zoo.pipeline.inference import InferenceModel
from zoo.common.nncontext import init_nncontext
from zoo.feature.image import *
from zoo.pipeline.nnframes import *

sc = init_nncontext("OpenVINO Python resnet_v1_50 Inference Example")
# pre-processing
infer_transformer = ChainedPreprocessing([ImageBytesToMat(),
                                         ImageResize(256, 256),
                                         ImageCenterCrop(224, 224),
                                         ImageMatToTensor(format="NHWC", to_RGB=True) ])

images = ImageSet.read(img_path, sc).\
    transform(infer_transformer).get_image().collect()
# load openvino model
model = InferenceModel()
model.load_openvino(model_path, weight_path)

# predict
for batch in images:
    model.predict(batch)
```

End-to-End Big Data and AI Pipelines

Seamless Scaling from Laptop to Production



Unified Analytics + AI Platform

Distributed TensorFlow*, Keras*, PyTorch* & BigDL on Apache Spark*

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





LEGAL NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit <http://www.intel.com/performance>.

Intel, the Intel logo, Xeon, Xeon phi, Lake Crest, etc. are trademarks of Intel Corporation in the U.S. and/or other countries.

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

© 2019 Intel Corporation