# ANALYTICS ZOO 统一的大数据分析 + AI 平台

*Analytics Zoo: A Unified Data Analytics + AI Platform*

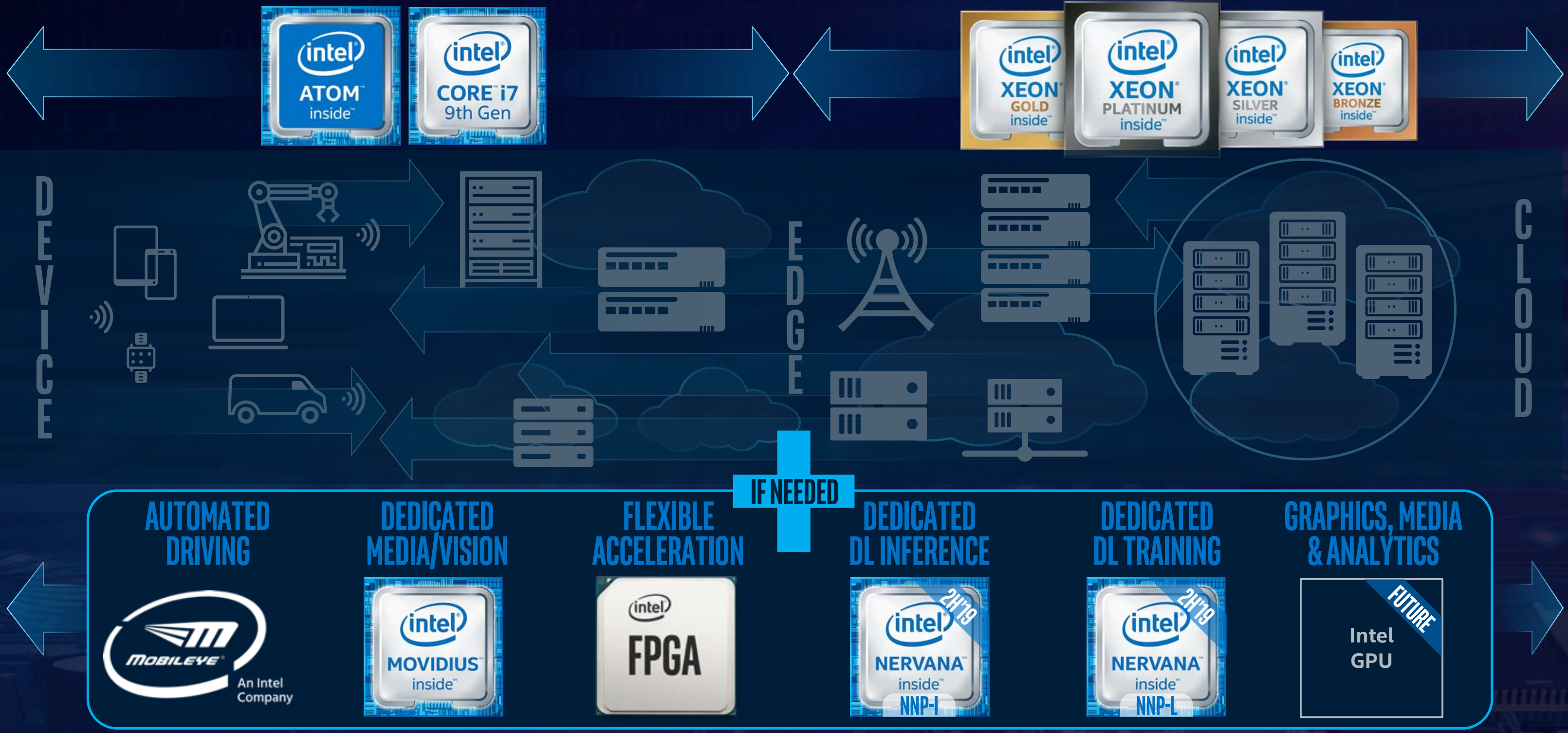
Li, Zhichao Wang, yang Huang, kai

# Agenda

- Analytics Zoo介绍
- 腾讯云: Sparkling 简单使用
- 动手实践:
  - 迁移学习(猫vs狗)
  - 文本分类
  - 推荐(Wide & Deep)
  - 图片切割(车)
  - VAE (可选)
  - 异常检测 (可选)

https://github.com/intel-analytics/analytics-zoo https://github.com/intel-analytics/ait2019

# One Size Does Not Fit All



AUTOMATED DRIVING

DEDICATED MEDIA/VISION

FLEXIBLE ACCELERATION

DEDICATED DL INFERENCE

DEDICATED DL TRAINING

GRAPHICS, MEDIA & ANALYTICS

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

# Speed Up Development

## Using Open AI Software

MACHINE LEARNING

DEEP LEARNING

### TOOLKITS
App developers

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

**OpenVINO™**
Deep learning inference deployment on CPU/GPU/FPGA/VPU for Caffe*, TensorFlow*, MXNet*, ONNX*, Kaldi*

**NAUTA**
Open source, scalable, and extensible distributed deep learning platform built on Kubernetes (BETA)

### LIBRARIES
Data scientists

**Python**
- Scikit-learn
- Pandas
- NumPy

**R**
- Cart
- Random Forest
- e1071

**Distributed**
- MlLib (on Spark)
- Mahout

**Intel-optimized Frameworks**

TensorFlow* · Caffe2* · ONNX* · mxnet* · PyTorch* · BigDL

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*

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

[1] An open source version is available at: 01.org/openvinotoolkit

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

Developer personas show above represent the primary user base for each row, but are not mutually-exclusive

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

# Real-World ML/DL Applications Are Complex Data Analytics Pipelines

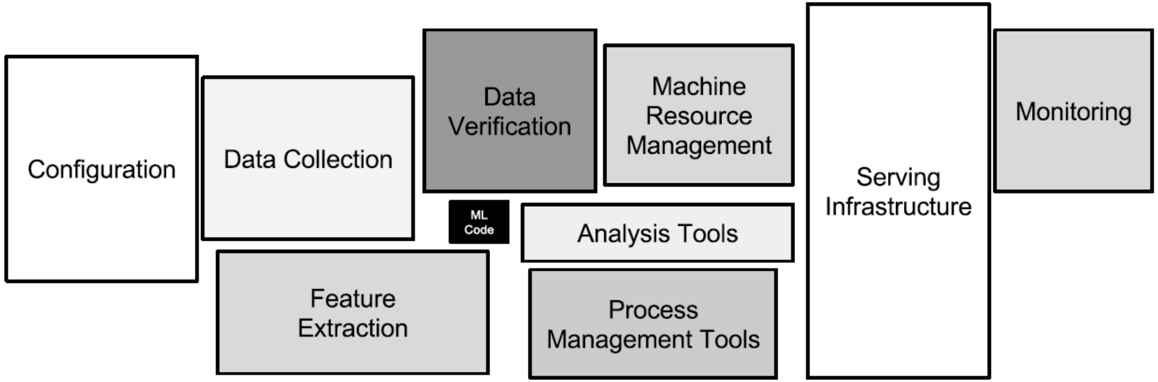
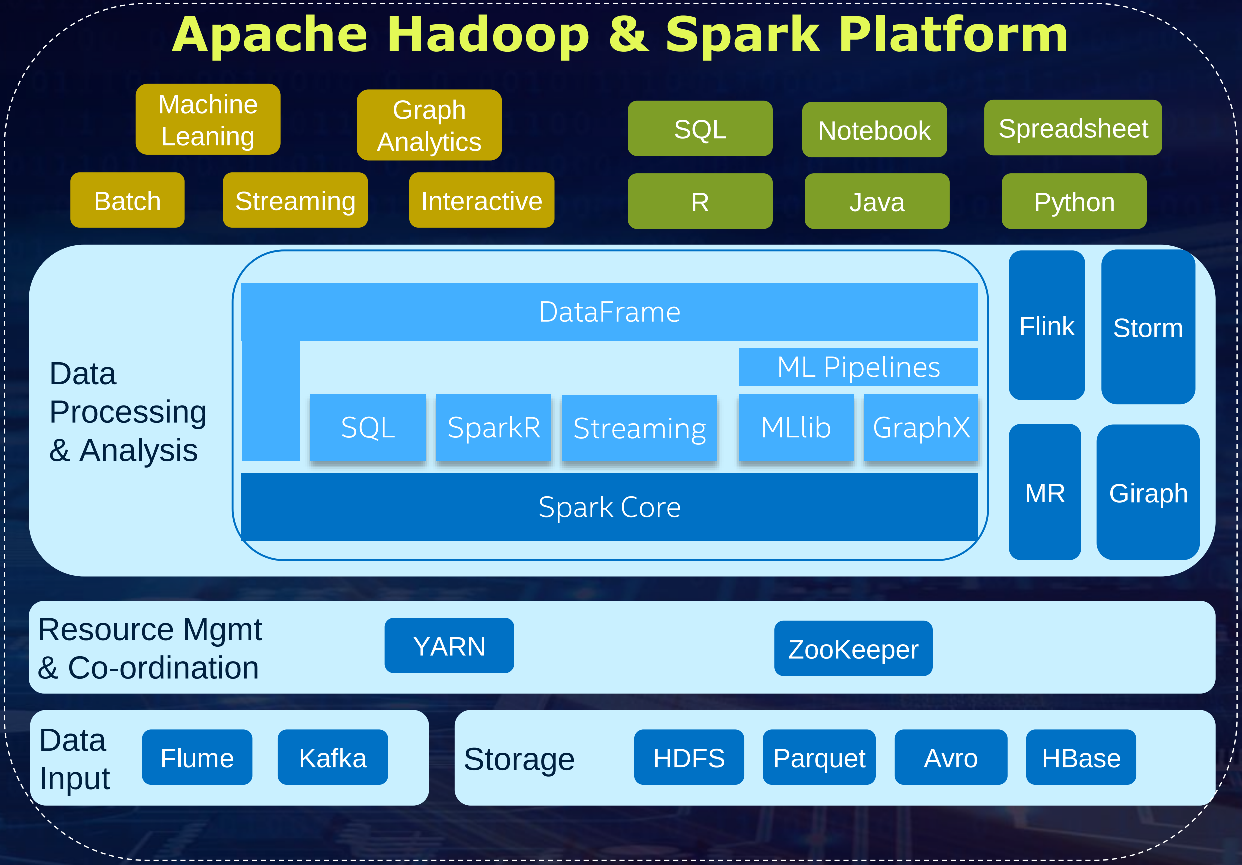


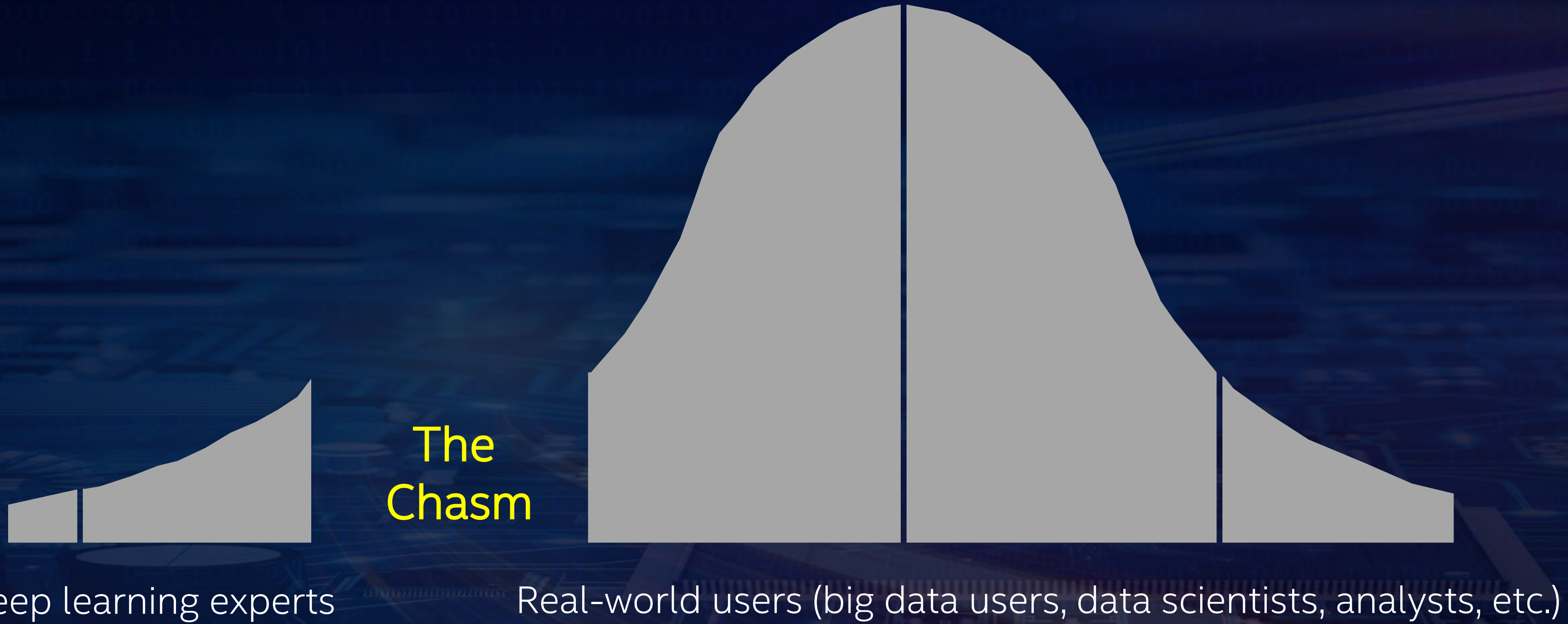
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

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

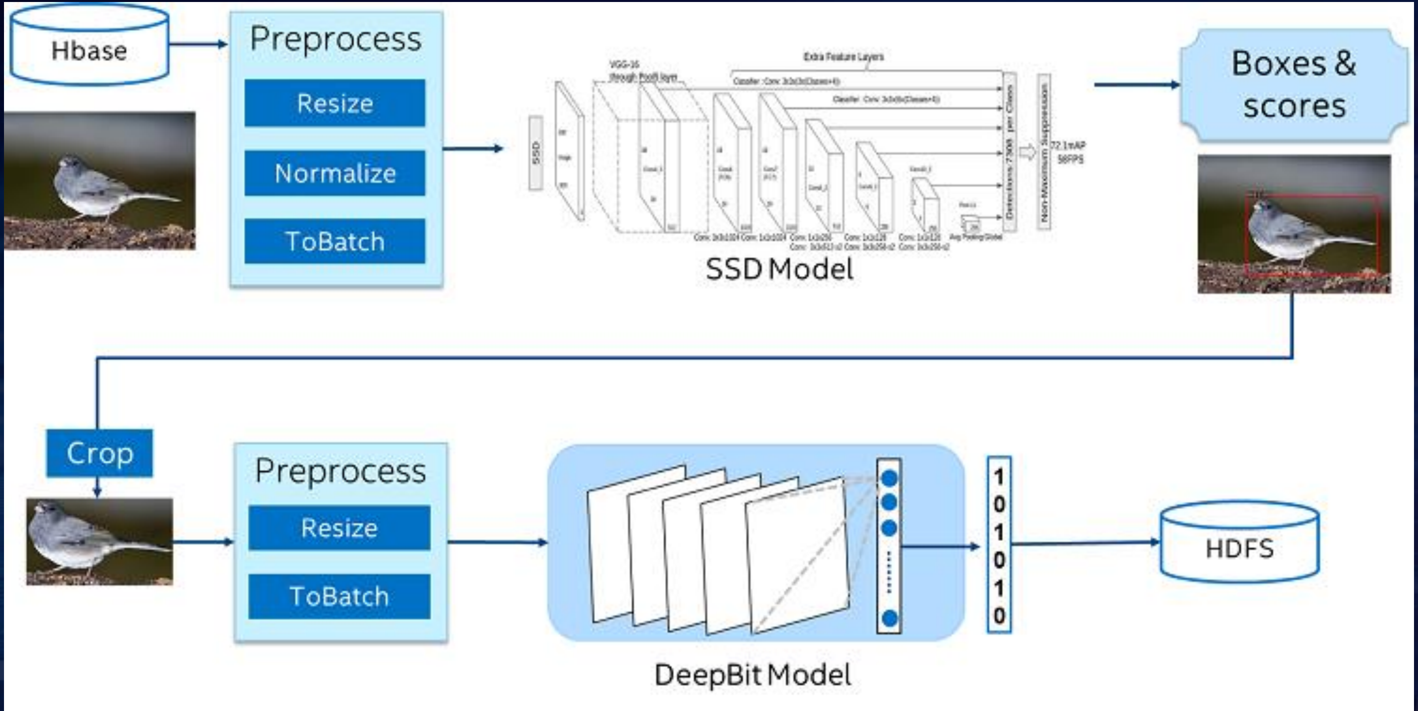
# Unified Big Data Analytics Platform



# Chasm b/w Deep Learning and Big Data Communities



# Large-Scale Image Recognition at JD.com



https://software.intel.com/en-us/articles/building-large-scale-image-feature-extraction-with-bigdl-at-jdcom

# AI on APACHE Spark™

BigDL

Distributed, High-Performance

**Deep Learning Framework**

for Apache Spark*

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

ANALYTICS ZOO

**Analytics + AI Platform**

Distributed TensorFlow*, Keras* and BigDL on Apache Spark*

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

## Unifying Analytics + AI on Apache Spark*

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

# AI on APACHE Spark™

## BigDL

Distributed, High-Performance

Deep Learning Framework

for Apache Spark*

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

## ANALYTICS ZOO

Analytics + AI Platform

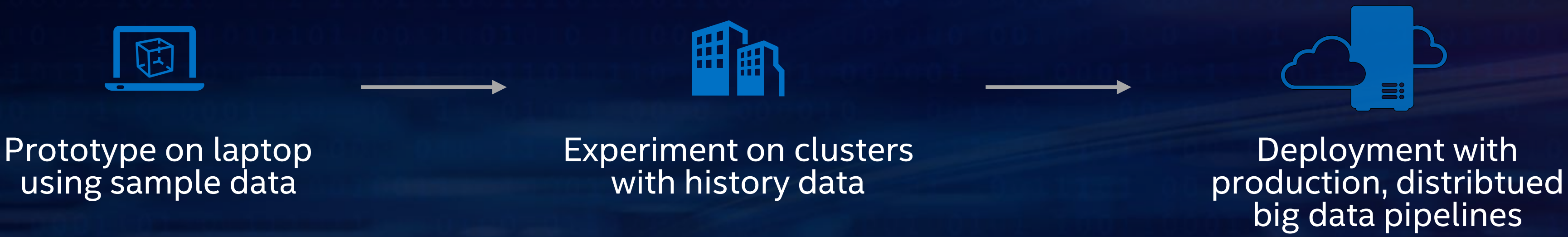
Distributed TensorFlow*, Keras* and BigDL on Apache Spark*

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

Unifying Analytics + AI on Apache Spark*

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

# Analytics Zoo: End-to-End DL Pipeline Made Easy for Big Data



Prototype on laptop using sample data

Experiment on clusters with history data

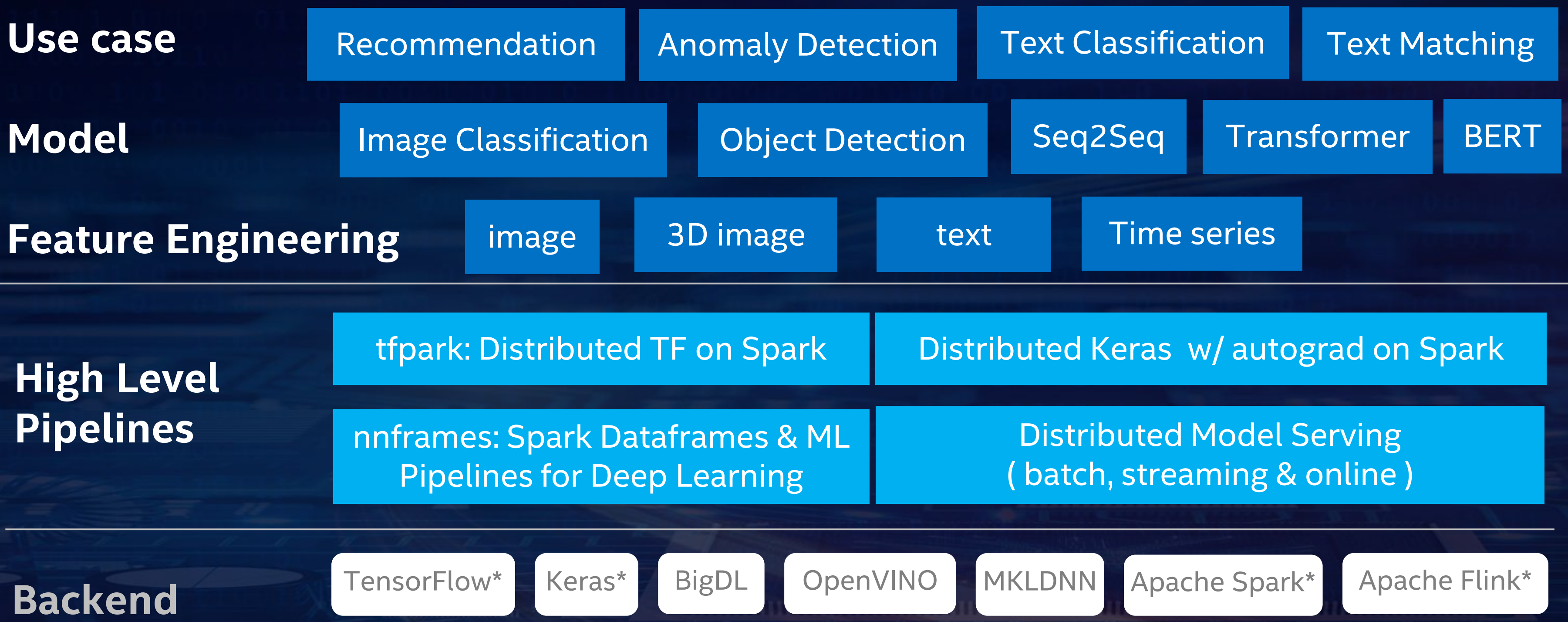
Deployment with production, distribtued big data pipelines

- “Zero” code change from laptop to distributed cluster
- Directly accessing production big data (Hadoop/Hive/HBase)
- Easily prototyping the end-to-end pipeline
- Seamlessly deployed on production big data clusters

# What is Analytics Zoo?

# Analytics Zoo

## Unified Analytics + AI Platform for Big Data



**Use case**: Recommendation | Anomaly Detection | Text Classification | Text Matching

**Model**: Image Classification | Object Detection | Seq2Seq | Transformer | BERT

**Feature Engineering**: image | 3D image | text | Time series

**High Level Pipelines**:

- tfpark: Distributed TF on Spark
- Distributed Keras w/ autograd on Spark
- nnframes: Spark Dataframes & ML Pipelines for Deep Learning
- Distributed Model Serving ( batch, streaming & online )

**Backend**: TensorFlow* | Keras* | BigDL | OpenVINO | MKLDNN | Apache Spark* | Apache Flink*

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

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

# Analytics Zoo

## Unified Analytics + AI Platform for Big Data

### Build end-to-end deep learning applications for big data

- Distributed *TensorFlow* on Spark
- *Keras* API (with autograd & transfer learning support) on Spark
- *nnframes*: native DL support for Spark DataFrames and ML Pipelines

### Productionize deep learning applications for big data at scale

- Plain Java/Python *model serving* APIs (w/ OpenVINO support)
- Support Web Services, Spark, Flink, Storm, Kafka, etc.

### Out-of-the-box solutions

- Built-in deep learning *models*, *feature engineering* operations, and reference *use cases*

# Distributed TF & Keras on Spark

- Data wrangling and analysis using PySpark

- Deep learning model development using TensorFlow or Keras

- Distributed training / inference on Spark

**Write TensorFlow code inline in PySpark program**

```
#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))
```

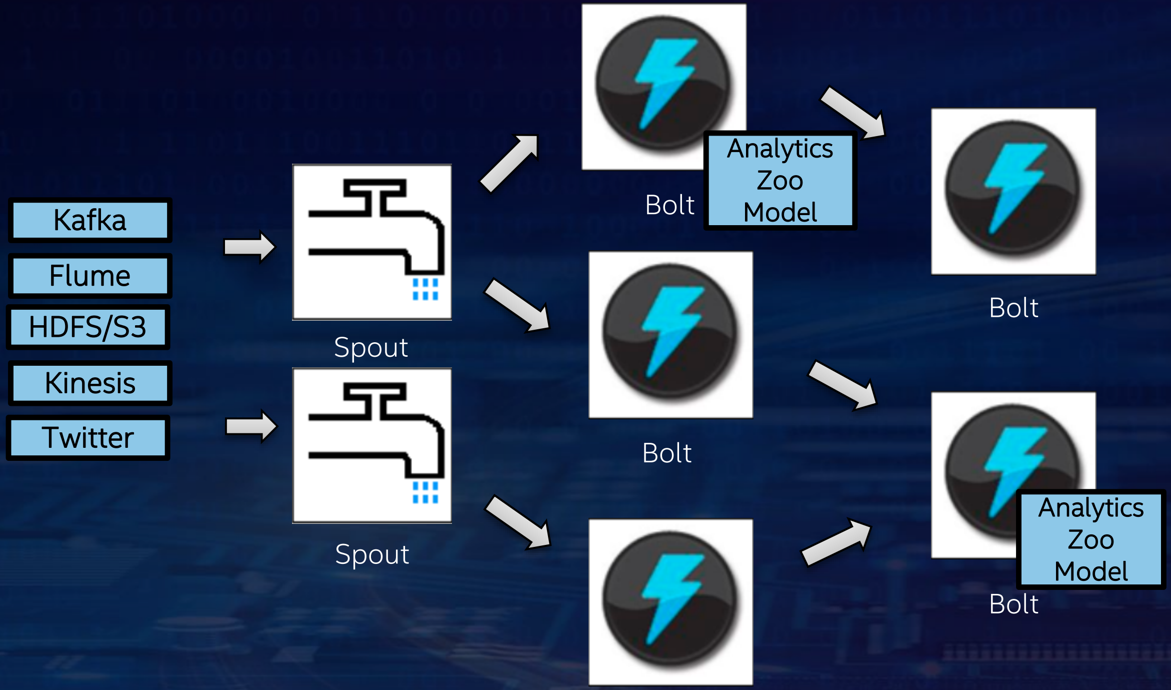
# Spark Dataframe & ML Pipeline for DL

```
#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
Estimater = NNEstimater(model, CrossEntropyCriterion()) \
                .setLearningRate(0.003).setBatchSize(40).setMaxEpoch(5) \
                .setFeaturesCol("image")
nnModel = estimater.fit(train_df)
```

# Distributed Model Serving

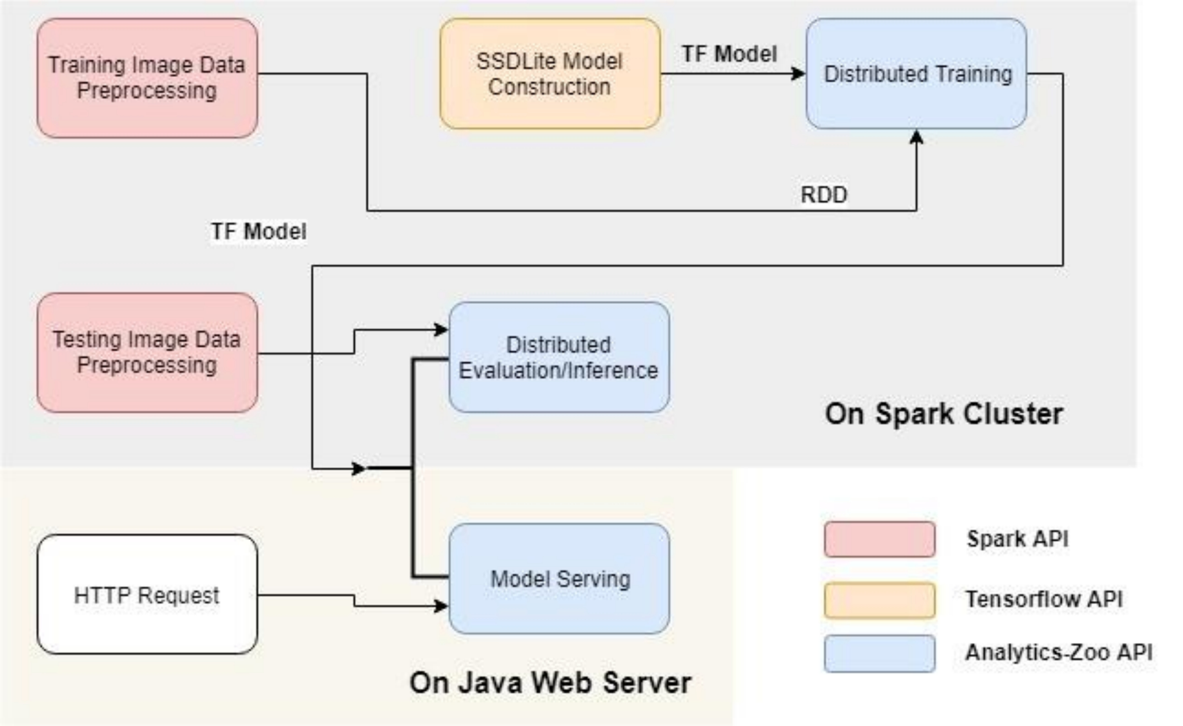


**Distributed model serving in Web Service, Flink, Kafka, Storm, etc.**

- **Plain Java or Python API, with OpenVINO and DL Boost (VNNI) support**

# Analytics Zoo Use Cases

# Computer Vision Based Product Defect Detection in Midea



Training Image Data Preprocessing

SSDLite Model Construction

TF Model

Distributed Training

RDD

TF Model

Testing Image Data Preprocessing

Distributed Evaluation/Inference

On Spark Cluster

HTTP Request

Model Serving

On Java Web Server

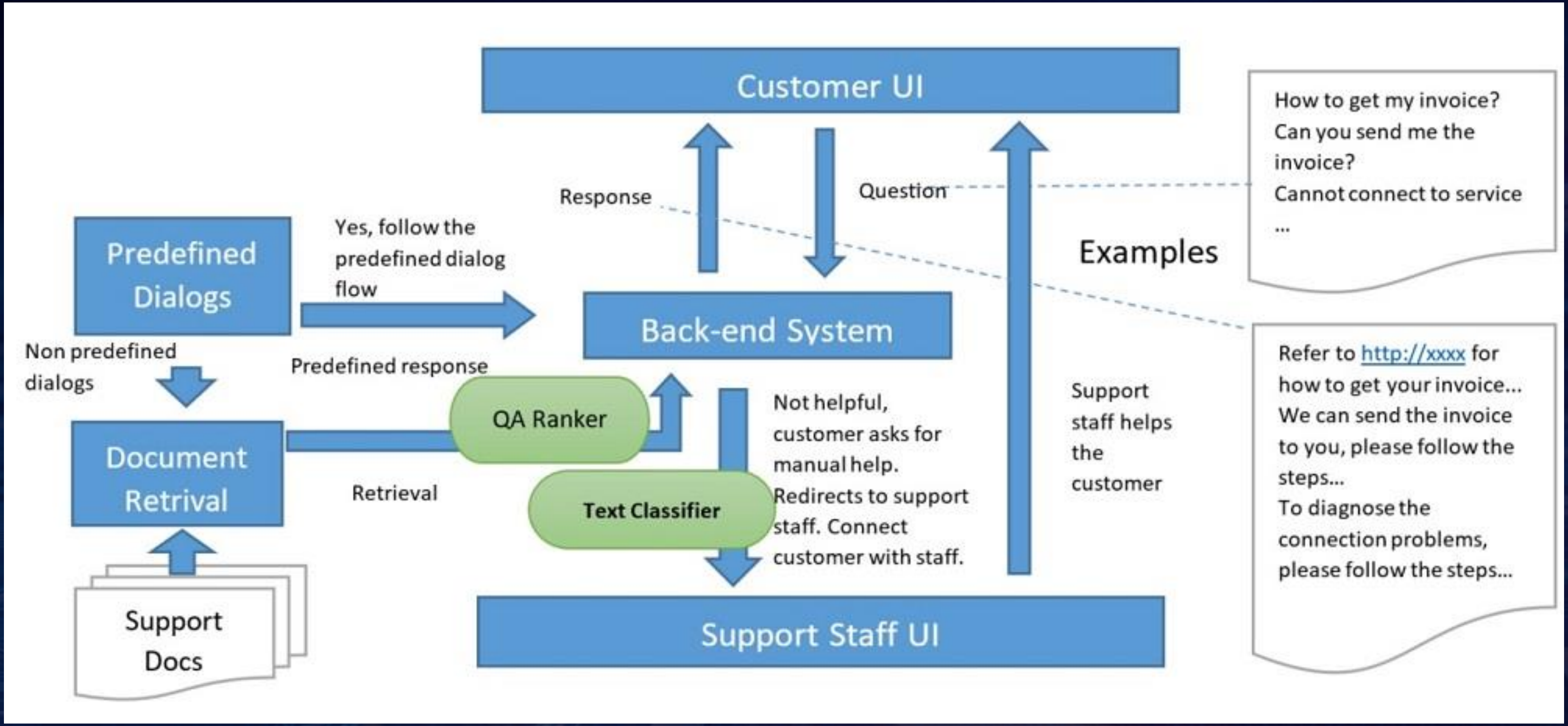
🟥 Spark API

🟧 Tensorflow API

🟦 Analytics-Zoo API

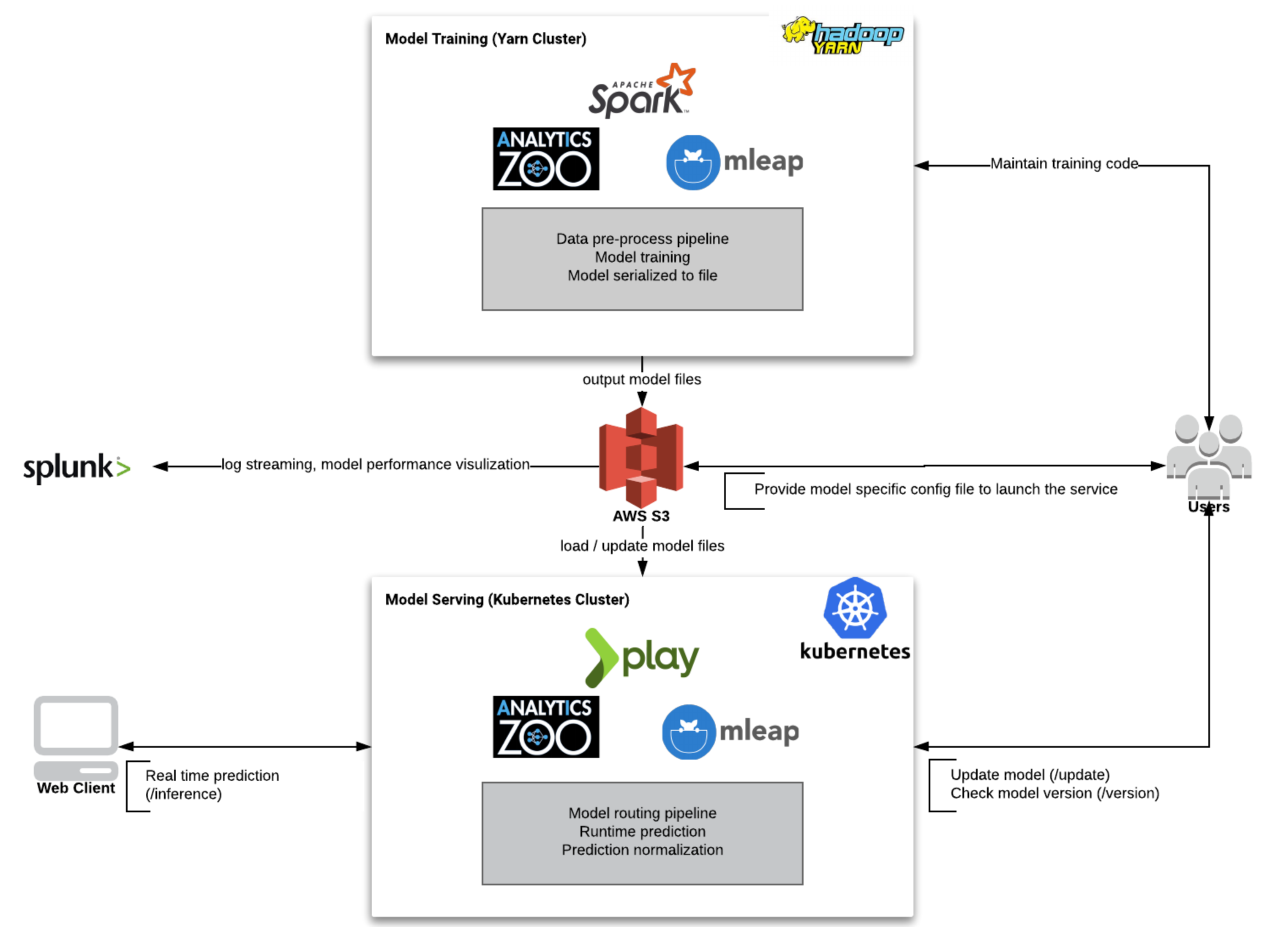
https://software.intel.com/en-us/articles/industrial-inspection-platform-in-midea-and-kuka-using-distributed-tensorflow-on-analytics

# NLP Based Customer Service Chatbot for Microsoft Azure


Customer UI
Response
Question
Examples
How to get my invoice?
Can you send me the invoice?
Cannot connect to service
...
Predefined Dialogs
Yes, follow the predefined dialog flow
Back-end System
Non predefined dialogs
Predefined response
QA Ranker
Document Retrival
Retrieval
Text Classifier
Not helpful, customer asks for manual help. Redirects to support staff. Connect customer with staff.
Support staff helps the customer
Refer to http://xxxx for how to get your invoice...
We can send the invoice to you, please follow the steps...
To diagnose the connection problems, please follow the steps...
Support Docs
Support Staff UI

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/

# Product Recommendations in Office Depot



Model Training (Yarn Cluster)

Data pre-process pipeline
Model training
Model serialized to file

output model files

AWS S3

log streaming, model performance visulization

Maintain training code

Provide model specific config file to launch the service

Users

load / update model files

Model Serving (Kubernetes Cluster)

Model routing pipeline
Runtime prediction
Prediction normalization

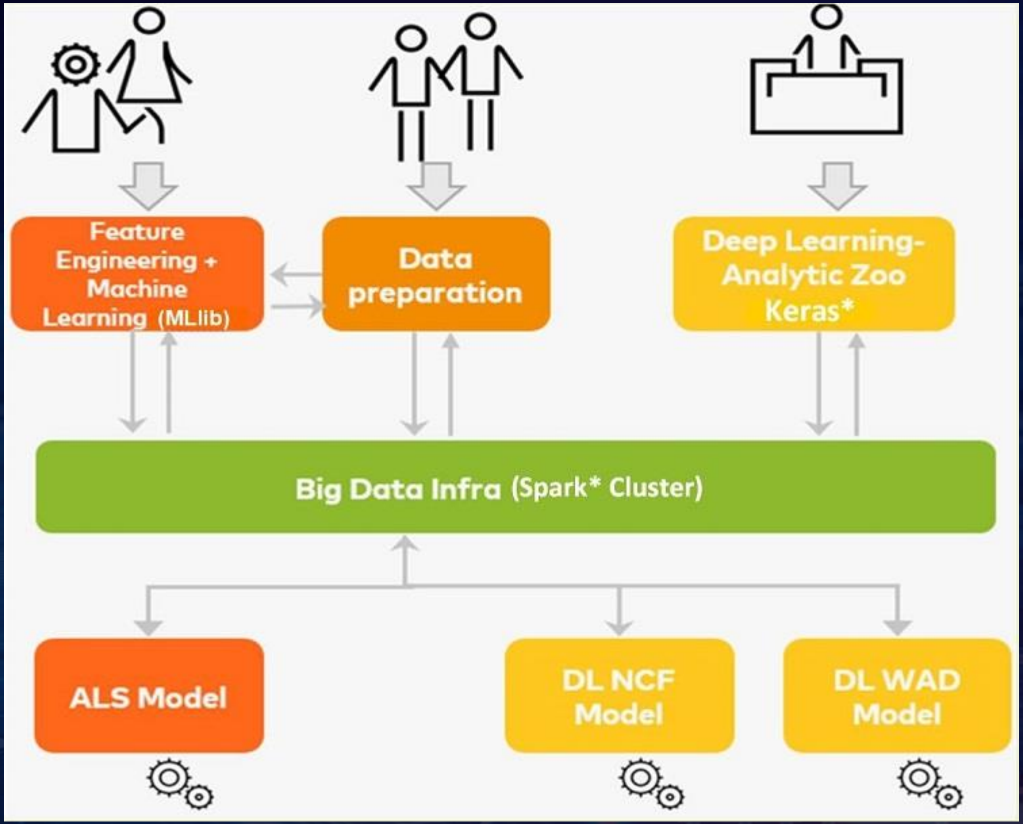
Web Client

Real time prediction
(/inference)

Update model (/update)
Check model version (/version)

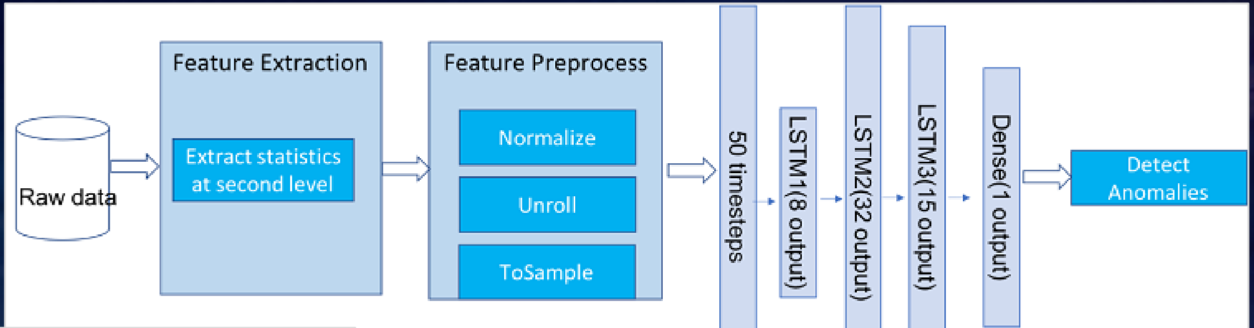
https://conferences.oreilly.com/strata/strata-ca-2019/public/schedule/detail/73079

# Recommender AI Service in MasterCard


Feature Engineering + Machine Learning (MLlib)
Data preparation
Deep Learning-Analytic Zoo Keras*
Big Data Infra (Spark* Cluster)
ALS Model
DL NCF Model
DL WAD Model

https://software.intel.com/en-us/articles/deep-learning-with-analytic-zoo-optimizes-mastercard-recommender-ai-service

# LSTM-Based Time Series Anomaly Detection for Baosight



Raw data

Feature Extraction

Extract statistics at second level

Feature Preprocess

Normalize

Unroll

ToSample

50 timesteps

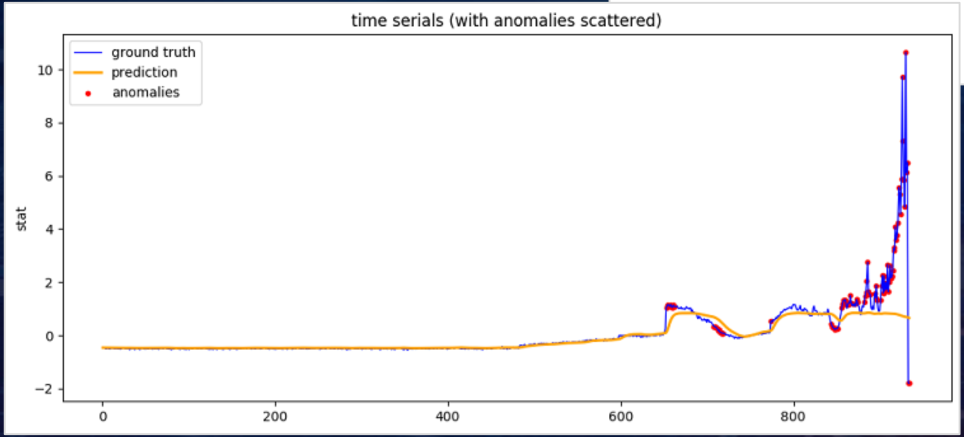
LSTM1(8 output)

LSTM2(32 output)

LSTM3(15 output)

Dense(1 output)

Detect Anomalies



time serials (with anomalies scattered)

https://software.intel.com/en-us/articles/lstm-based-time-series-anomaly-detection-using-analytics-zoo-for-apache-spark-and-bigdl

# And Many More

## TECHNOLOGY

bluedata®

cloudera®

CRAY
THE SUPERCOMPUTER COMPANY

databricks

DELL EMC

GIGASPACES
innovate with confidence

Lightbend

Qubole

## CLOUD SERVICE PROVIDERS

Alibaba Cloud
aliyun.com

aws

Cloud Dataproc

Azure

IBM Cloud

Telefonica

KINGSOFT

## END USERS

cdhi

中国电信
CHINA TELECOM

JD.COM 京东

Midea®

UnionPay
银联

http://software.intel.com/bigdl/build

*Not a full list*

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

# 本次课程平台基于腾讯云SPARKLING数据仓库

- Sparkling 云上一站式大数据解决方案
  - 产品信息: https://cloud.tencent.com/product/sparkling

- 欢迎参加我们的Talk和Booth
  - Talk
    - 报告厅，周五 13:10
    - Sparkling: 基于Apache Spark进行一站式机器学习
  - Booth:
    - 会议中心2层，周四、周五
    - 腾讯云Sparkling + Intel Analytics Zoo云上数仓的数据科学方案

# Deep Learning Made Easy for Big Data

ANALYTICS ZOO

Unified Analytics + AI Platform

Distributed TensorFlow*, Keras* and BigDL on Apache Spark*

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

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

知未来 创未来

# LEGAL 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.