

Analytics-Zoo: 统一的大数据分析+AI平台

Analytics Zoo: A Unified Data Analytics + AI Platform

利智超

Intel 软件架构师

大数据分析和人工智能创新院

Why Analytics-Zoo



Distributed, High-Performance
Deep Learning Framework
for Apache Spark*

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



Analytics + AI Platform
Distributed TensorFlow*, Keras* and
BigDL on Apache Spark*

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

Unifying Analytics + AI on Apache Spark*

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

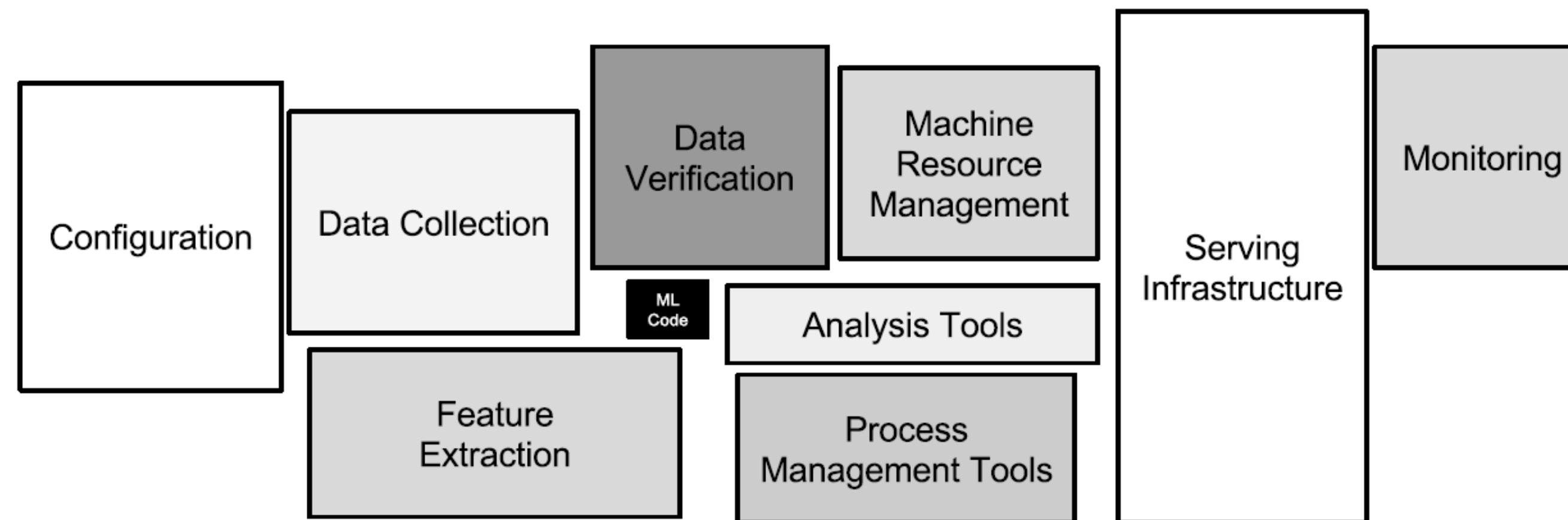
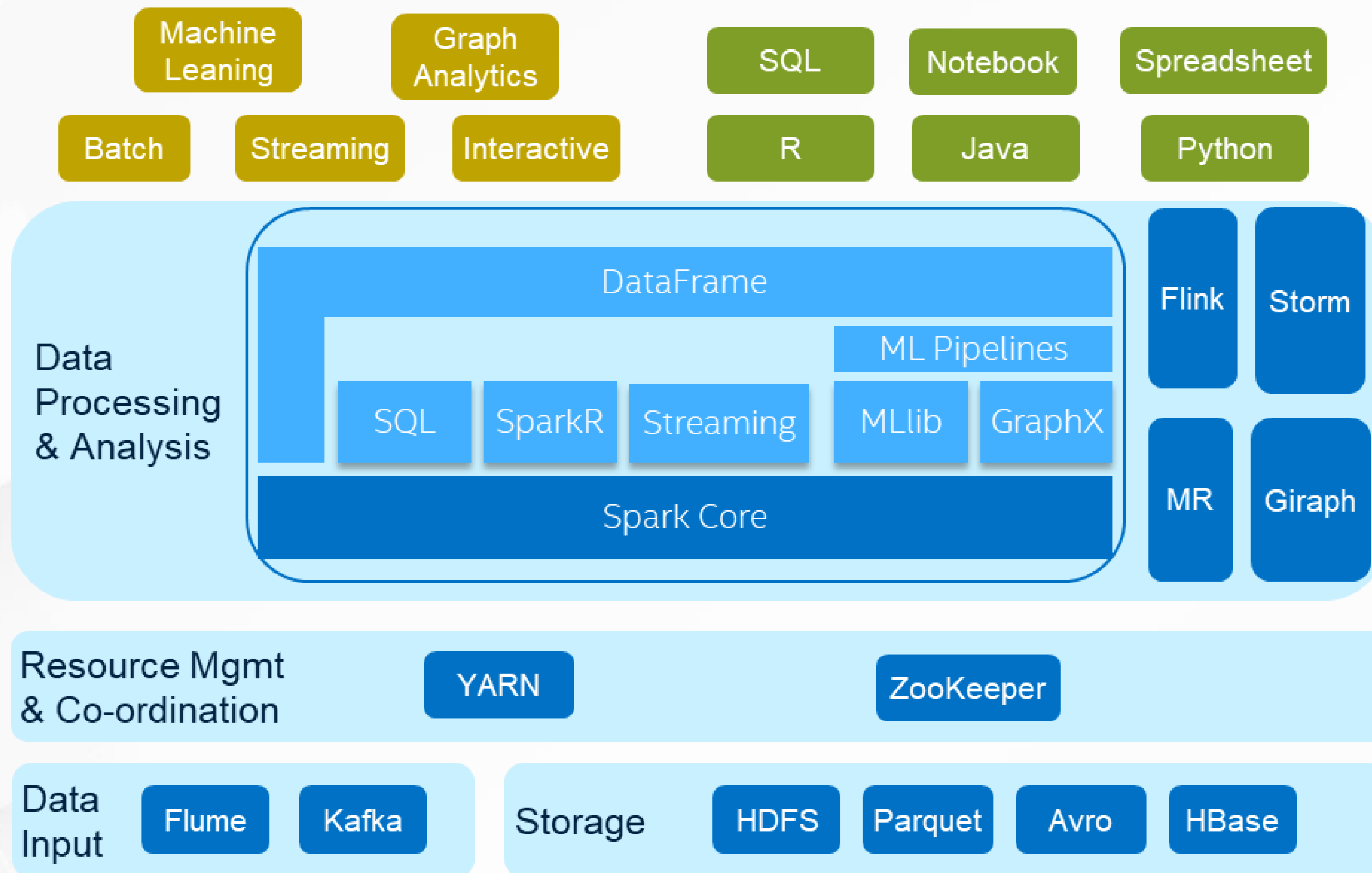


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

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

Unified Big Data Analytics Platform

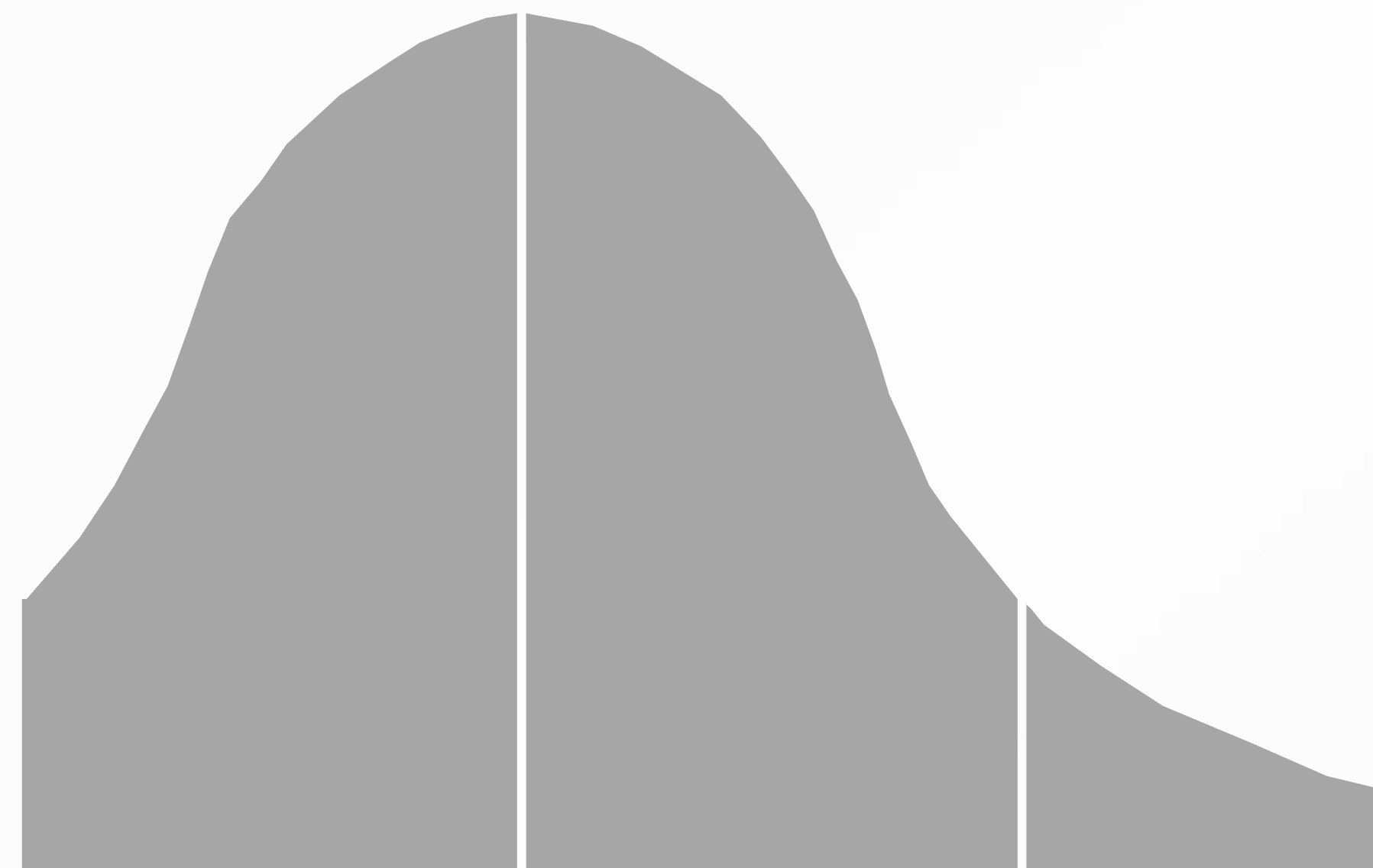


Chasm b/w Deep Learning and Big Data Communities



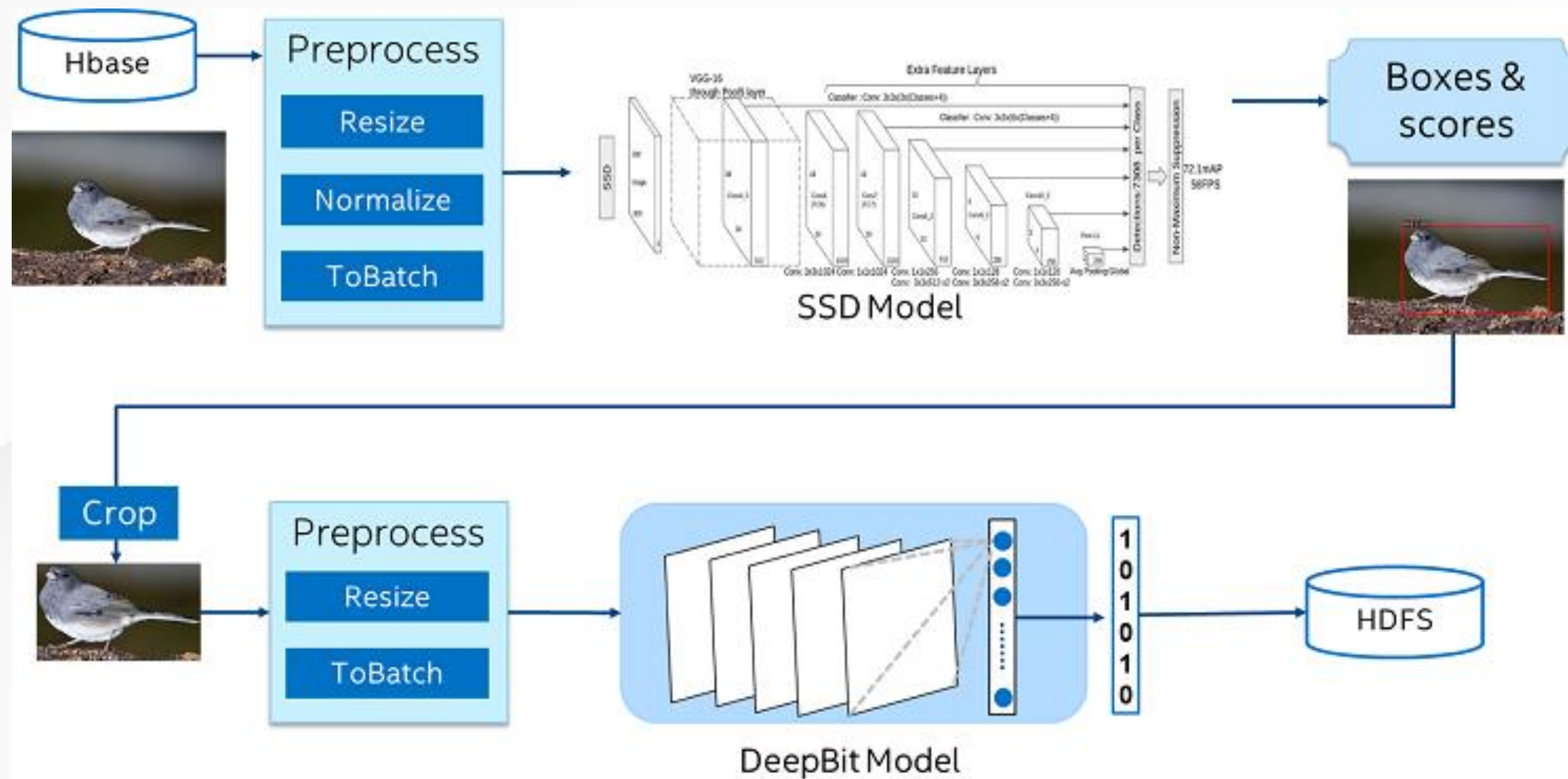
Deep learning experts

The
Chasm



Real-world users (big data users, data scientists, analysts, etc.)

Large-Scale Image Recognition



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



Standard Spark jobs

- No changes to the Spark or Hadoop clusters needed

Data parallel

- Each Spark task runs the same model on a subset of the data (batch)

“Zero” code change

- Directly support TensorFlow, Keras and Caffe Model

Seamlessly deployed on production big data clusters

- Only need to install on driver node.

What's Analytics-Zoo



Distributed, High-Performance
Deep Learning Framework
for Apache Spark*

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



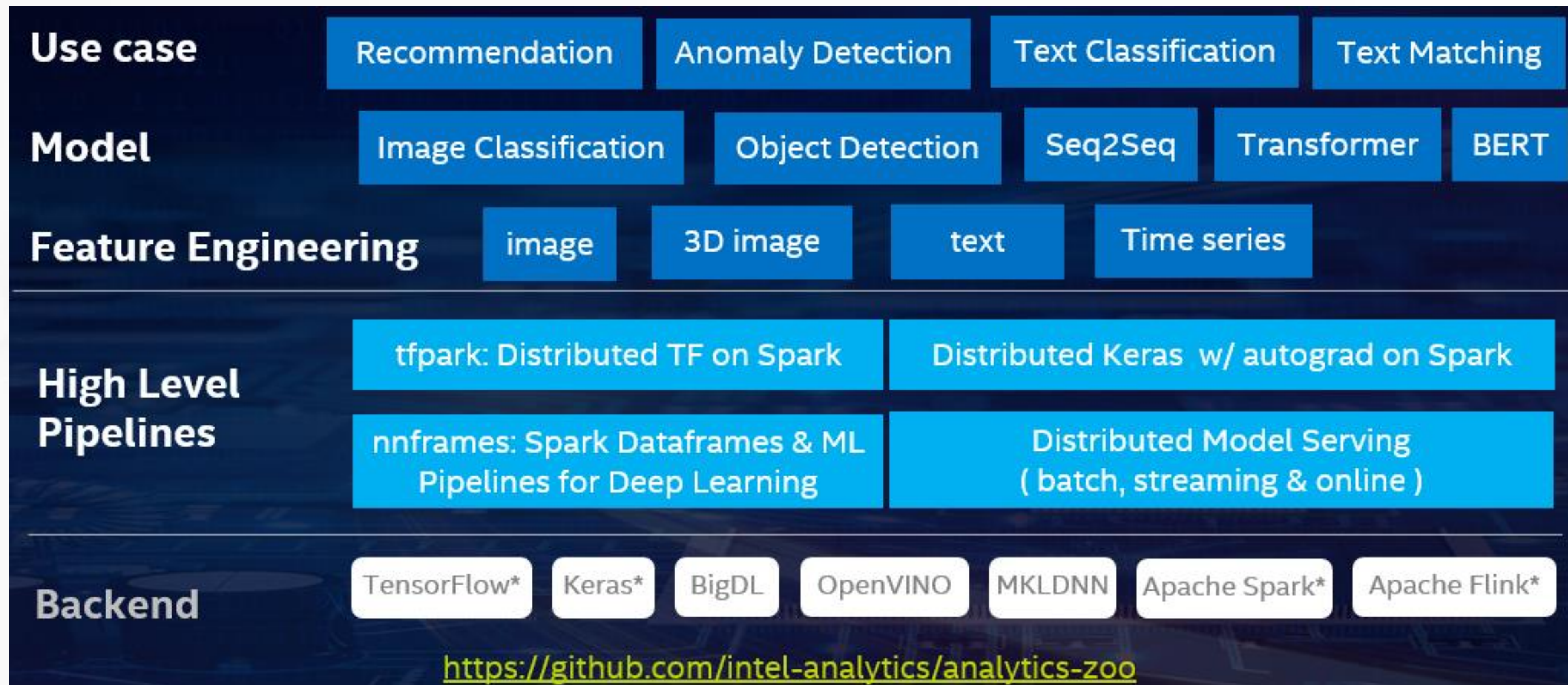
Analytics + AI Platform

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

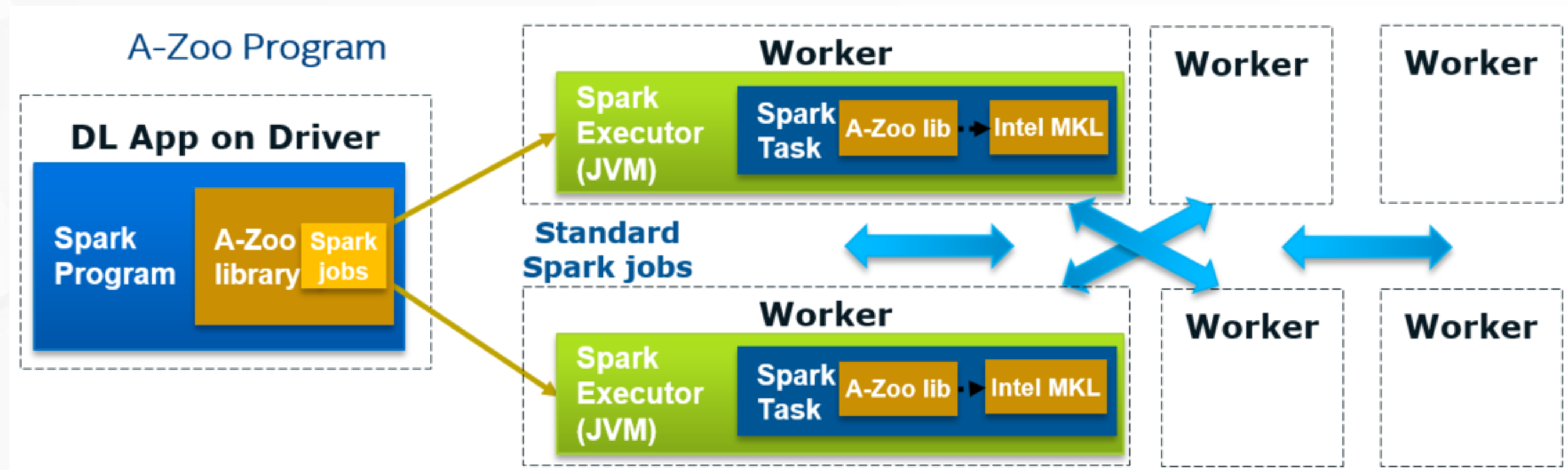
<https://github.com/intel-analytics/analytics-zoo>

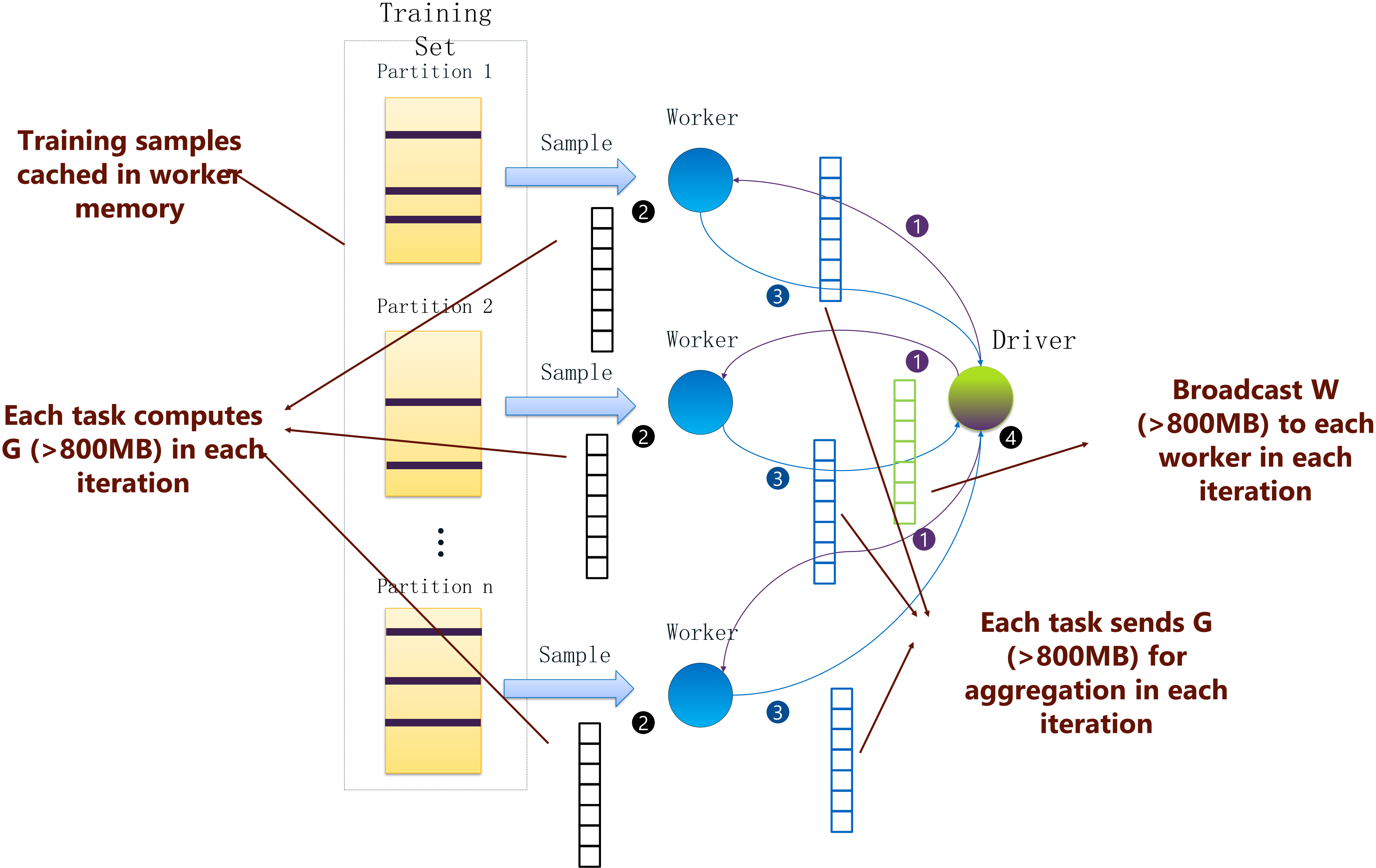
Unifying Analytics + AI on Apache Spark*

Analytics-Zoo: Unified Analytics + AI Platform for BigData

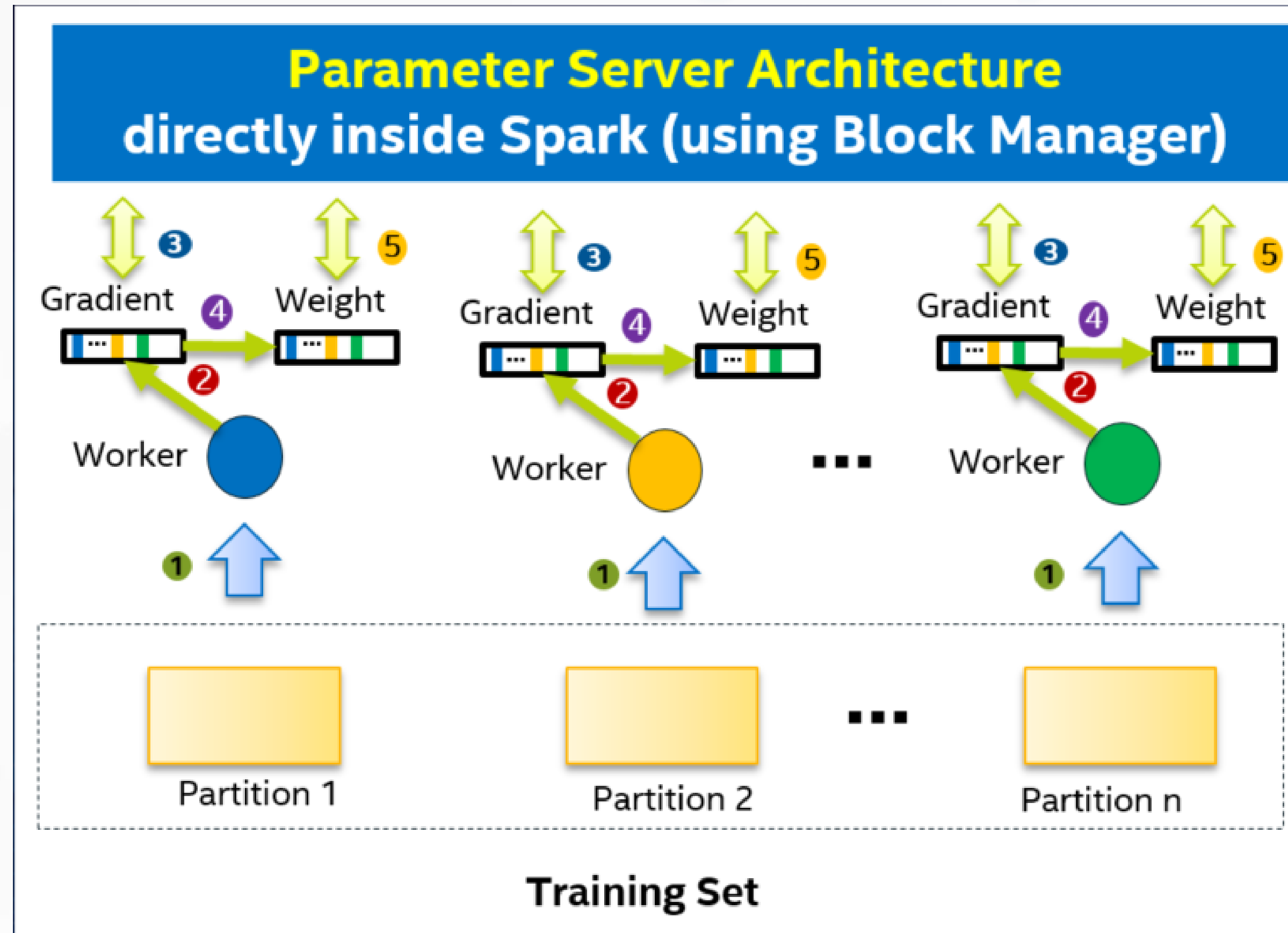


Analytics-Zoo: Run as Standard Spark Programs





Distributed Training in Analytics-Zoo



Peer-2-Peer All-Reduce synchronization

Distributed TF & Keras on Spark

Write TensorFlow code inline in PySpark program

- 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))
```

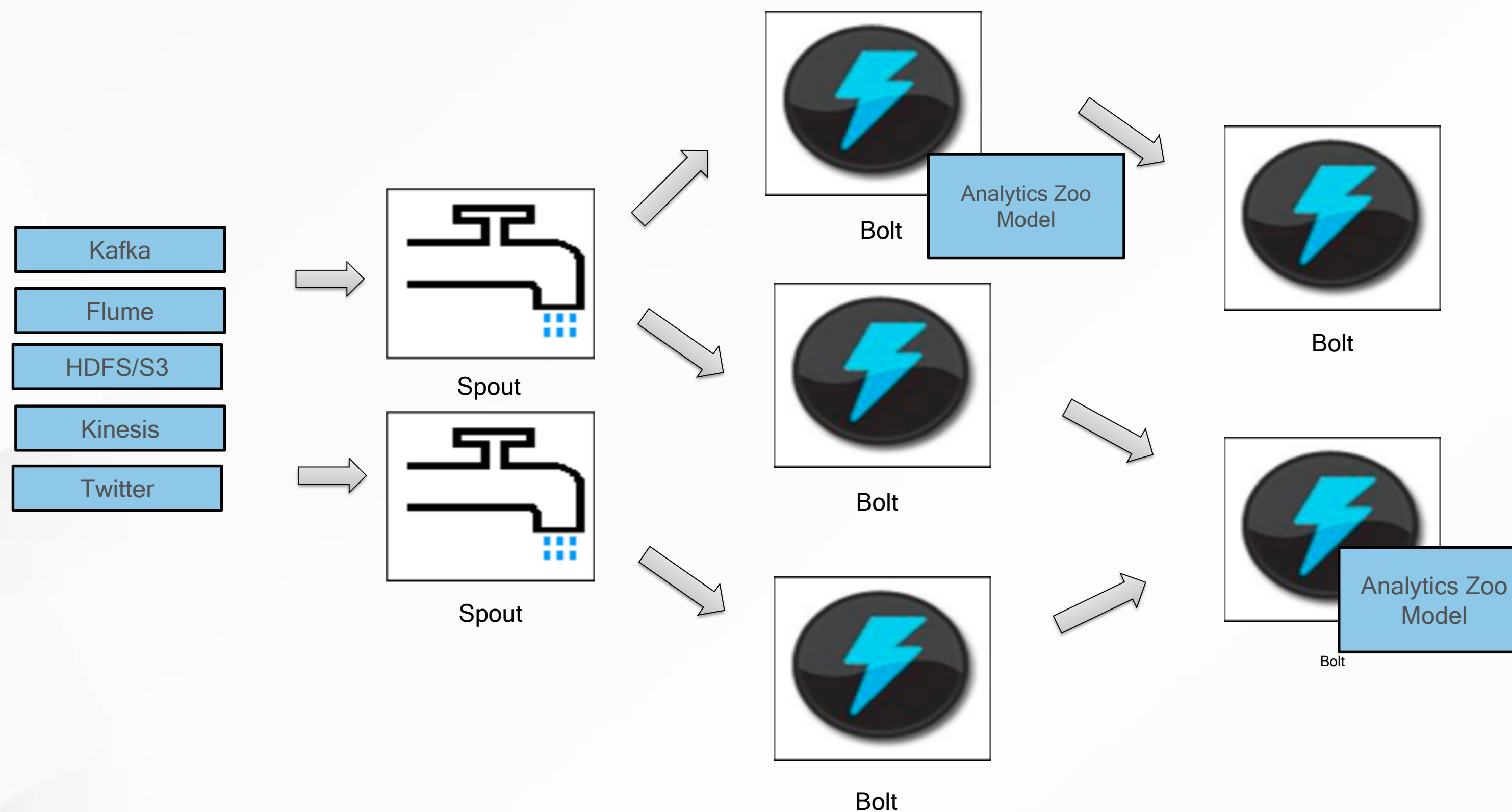

Spark Dataframe & ML Pipeline for DL

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

Distributed Model Serving

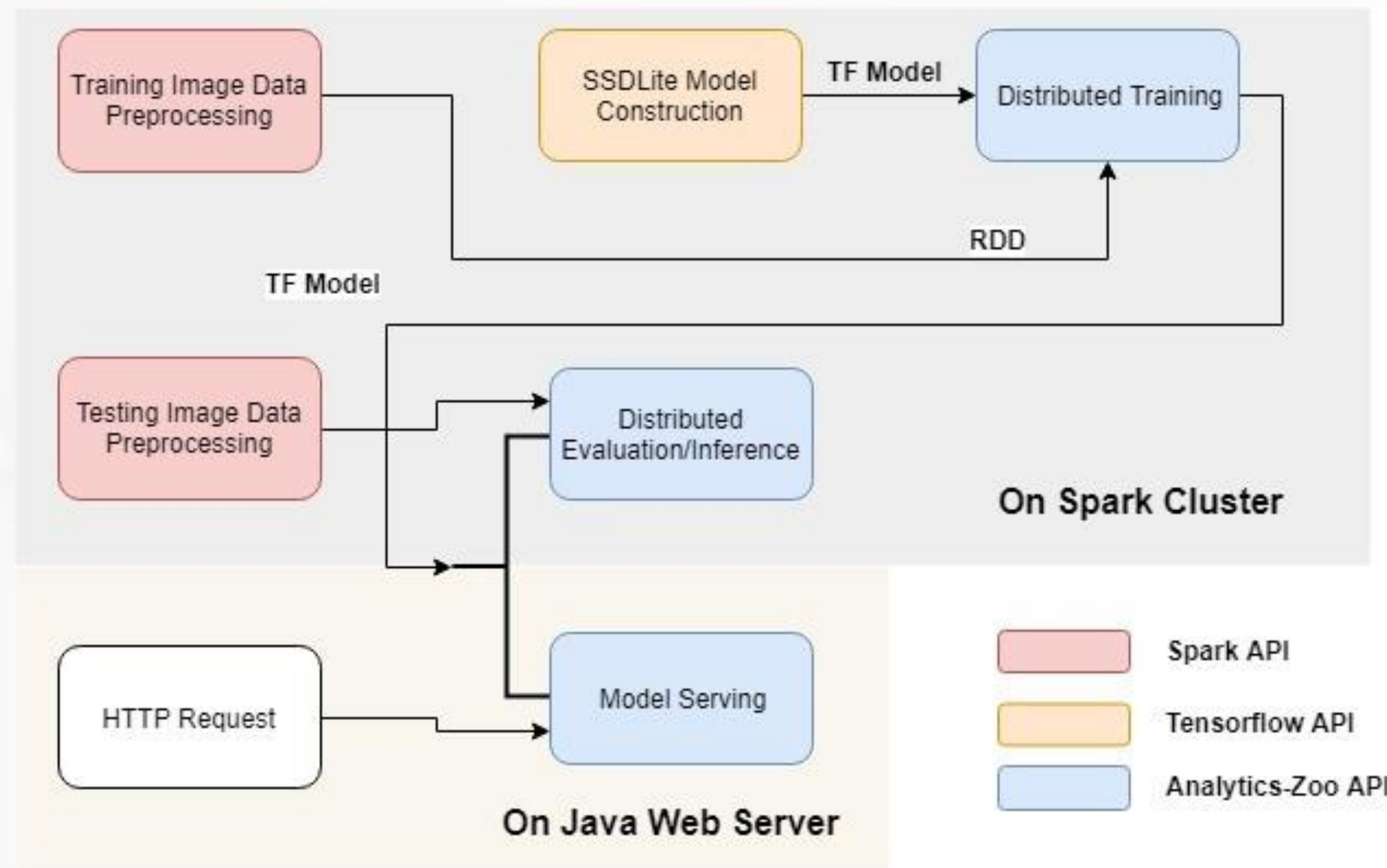


Distributed model serving in Web Service, Flink, Kafka, Storm, etc.

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

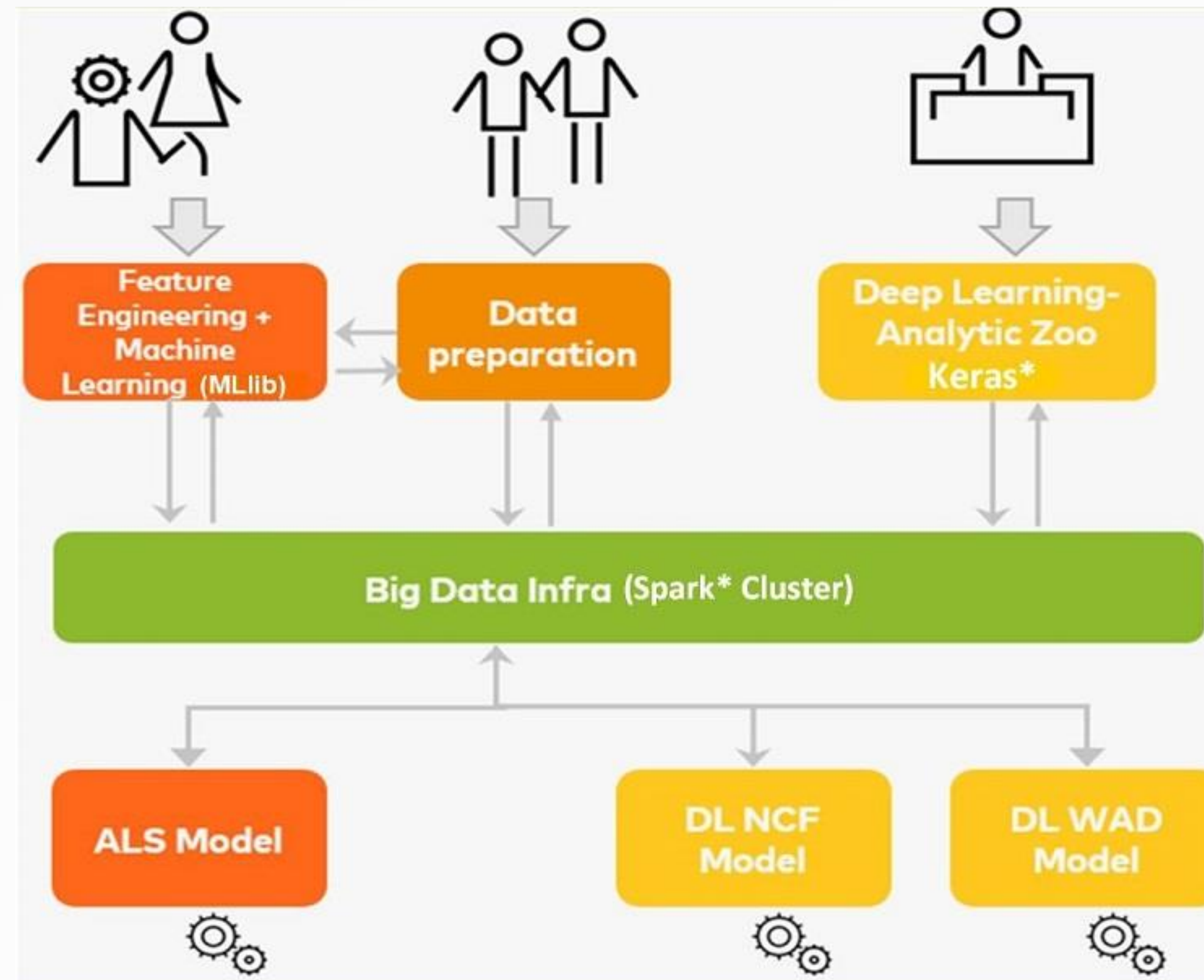
Analytics-Zoo use cases

Computer vision Based Product Defect Detection in Midea



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

Recommender AI Service in MasterCard



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

Deep Learning Made Easy for BigData



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.

© 2019 Intel Corporation



阿里云开发者社区

扫码加入社群
与志同道合的码友一起
Code Up

大数据计算开发者...



该群属于“阿里云ACE”部门群，仅组织内部成员可以加入，如果组织外部人员收到此分享，需要先申请加入该组织。

谢谢！