Building End-to-End, Integrated Big Data Analytics and AI Solution

Sajan Govindan
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

ANALYTICS ZOO

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

OpenVINO

NAUTA

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

DEEP LEARNING

LIBRARIES
Data scientists

PYTHON
• Scikit-learn
• Pandas
• NumPy

R
• Cart
• RandomForest
• g1071

DISTRIBUTED
• MLlib (on Spark)
• Mahout

INTEL-OPTIMIZED FRAMEWORKS

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

INTEL-OPTIMIZED FRAMEWORKS

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

KERNALS
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)

© 2019 Intel Corporation

Visit: www.intel.ai/technology

*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.

1. An open source version is available at 01.org/openvino/zoo/
Building End-to-End, Integrated Data Analytics & AI Solutions

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

[https://github.com/intel-analytics/bigdl](https://github.com/intel-analytics/bigdl)

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

[https://github.com/intel-analytics/analytics-zoo](https://github.com/intel-analytics/analytics-zoo)

*Other names and brands may be claimed as the property of others.*
Real-World ML/DL Solutions Are Complex Data Analytics Pipelines

## Analytics Zoo

### End-to-End, Integrated Data Analytics + AI Platform

<table>
<thead>
<tr>
<th>Use case</th>
<th>Recommendation</th>
<th>Anomaly Detection</th>
<th>Text Classification</th>
<th>Text Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Image Classification</td>
<td>Object Detection</td>
<td>Seq2Seq</td>
<td>Transformer</td>
</tr>
<tr>
<td>Feature Engineering</td>
<td>image</td>
<td>3D image</td>
<td>text</td>
<td>time series</td>
</tr>
<tr>
<td>High Level Pipelines</td>
<td>tfpark: Distributed TF on Big Data</td>
<td>Distributed Keras w/ autograd on Big Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>nnframes: Spark Dataframes &amp; ML Pipelines for Deep Learning</td>
<td>Distributed Model Serving (batch, streaming &amp; online)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backend/Library</td>
<td>TensorFlow</td>
<td>Keras</td>
<td>PyTorch</td>
<td>BigDL</td>
</tr>
<tr>
<td></td>
<td>Ray</td>
<td>MKLDNN</td>
<td>OpenVINO</td>
<td>Intel® Optane™ DCPMM</td>
</tr>
</tbody>
</table>

[https://github.com/intel-analytics/analytics-zoo](https://github.com/intel-analytics/analytics-zoo)
Distributed TF & Keras on Spark in Analytics Zoo

- 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**

```python
# 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))
```
Deep Learning Pipelines for High Energy Physics at CERN using Apache Spark and Analytics Zoo
THE NEES: Improve event selection accuracy at the particle detectors.

THE CHALLENGE: LHC is generating 1 Petabyte per second of data with particle collision events happening every 25 ns! Event filtering system accuracy improvement can provide large savings for data analysis resources (compute and storage).

THE SOLUTION: Use Analytics Zoo & Apache Spark on Intel® Xeon® Scalable servers to implement full data pipeline for training a topology classifier for event filtering at LHC.

THE RESULT: Successful test of end-to-end full data pipeline particle classifier implementation that easily scale out.

“Analytics Zoo & BigDL allowed us to Easily Scale-out Deep Learning Training on Apache Spark clusters running on Intel® Xeon® servers, and enabled our researchers to successfully develop an end-to-end data pipeline to improve real-time event selection at the Large Hadron Collider”

Maria Girone, Chief Technology Officer, CERN openlab

1 This CERN's solution is a proof of concept and not in production yet.

See proposed solution details at http://db-blog.web.cern.ch/blog/luca-canali/machine-learning-pipelines-high-energy-physics-using-apache-spark-bigdl

*Other names and brands may be claimed as the property of others.
Particle Classifier for High Energy Physics in CERN

Deep learning pipeline for physics data

Data Pipeline

- **Data Ingestion**
  - Read physics data and feature engineering

- **Feature Preparation**
  - Prepare input for Deep Learning network

- **Model Development**
  - 1. Specify model topology
  - 2. Tune model topology on small dataset

- **Training**
  - Train the best model

Leveraging Apache Spark and Analytics Zoo in Python Notebooks


And Many More

TECHNOLOGY
- bluedata
- cloudera
- CRAY
- databricks
- Inspur
- GIGASCAPES
- Lightbend
- Quable

CLOUD SERVICE PROVIDERS
- Alibaba Cloud (aliyun.com)
- AWS
- Azure
- Tencent
- Baidu
- IBM Cloud
- KINGSOFT

END USERS
- cdhi
- Telefonica
- China Telecom
- THE WORLD BANK
- JD.COM
- CERN openlab
- Midea
- CISCO
- UnionPay

software.intel.com/data-analytics

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