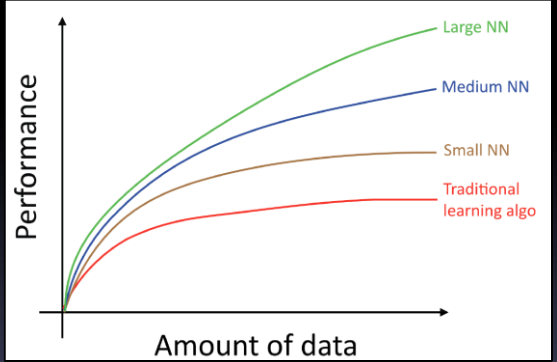
Intel data analytics solution at scale

Zhichao Li (zhichao.li@intel.com)

Motivations

Data Scale Driving Deep Learning Process



“Machine Learning Yearning”, Andrew Ng, 2016

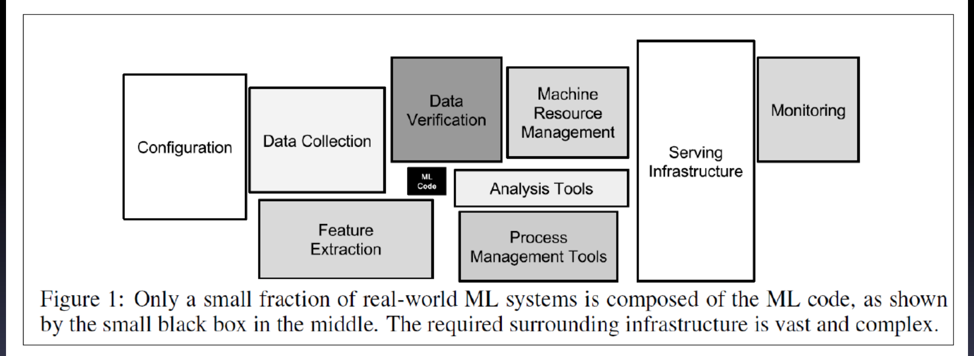
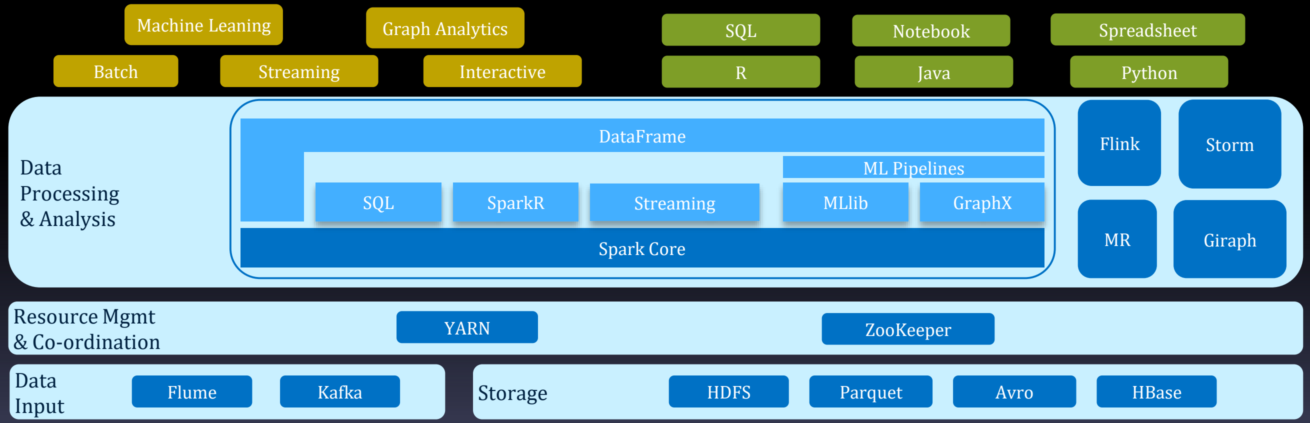


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

“Hidden Technical Debt in Machine Learning Systems”, Google, NIPS 2015 paper

Hadoop Becoming the Center of Data Gravity

Hadoop & Spark Platform

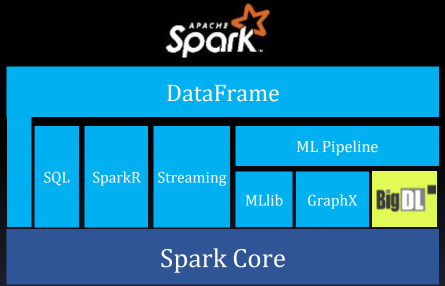


Overview

BigDL

Bringing Deep Learning To Big Data Platform

- Distributed deep learning framework for Apache Spark*
- Make deep learning more accessible to big data users
  data scientists
  - Write deep learning applications as ***standard Spark programs***
  - Run on existing Spark/Hadoop clusters ***no changes needed***
- Feature parity with popular deep learning frameworks
  - E.g., Caffe, Torch, Tensorflow, etc.
- High performance
  - Powered by Intel MKL and multi-threaded programming
- Efficient scale-out
  - Leveraging Spark for distributed training & inference



https://github.com/intel-analytics/BigDL
https://bigdl-project.github.io/

Analytics Zoo

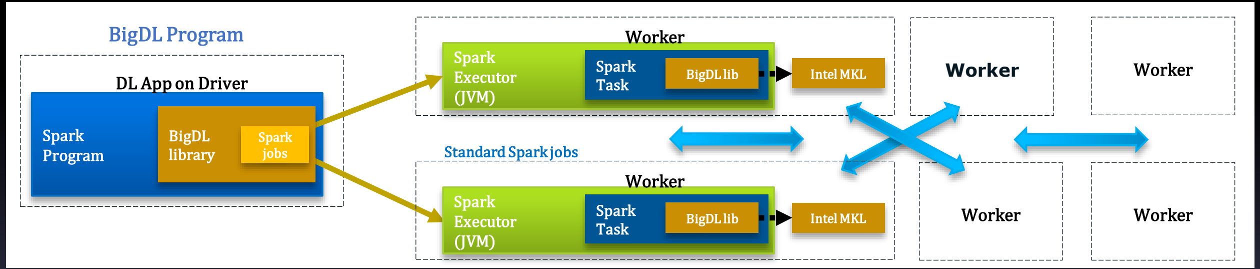
Build and Productionize Deep Learning Apps for Big Data at Scale

Reference Use Cases	• Anomaly detection • Sentiment analysis • Fraud detection • Chatbot, sequence prediction, etc.
Built-In Deep Learning Models	• Image classification • Object detection • Text classification • Recommendations • Sequence-to-sequence, GAN, etc.
Feature Engineering	Feature transformations for • Image, text, 3D imaging, time series, speech, etc.
High-Level Pipeline APIs	• Native deep learning support in Spark DataFrames and ML Pipelines • Autograd, Keras and transfer learning APIs for model definition • Support for model serving/inference pipelines
Backbends	Spark, BigDL, TensorFlow, etc.

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

https://analytics-zoo.github.io/

BigDL Run as Standard Spark Programs



Standard Spark jobs

- No changes to the Spark or Hadoop clusters needed

Iterative

- Each iteration of the training runs as a Spark job

Data parallel

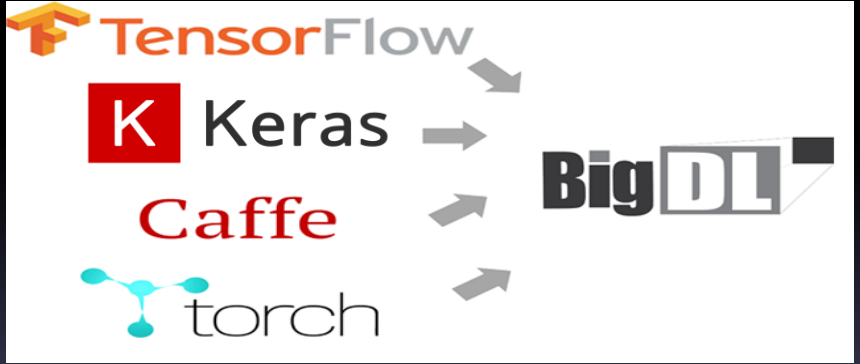
- Each Spark task runs the same model on a subset of the data (batch)

Models Interoperability Support

(e.g., between TensorFlow, Keras, Caffe, Torch, BigDL models)

Load existing TensorFlow, Keras, Caffe, Torch Model

- Useful for inference and model fine-tuning
- Allows for transition from single-node for distributed application deployment
- Allows for model sharing between data scientists and production engineers



Use Cases

Cloud & Big Data Platforms

Running BigDL, Deep Learning for Apache Spark, on AWS* (Amazon* Web Service)
https://aws.amazon.com/blogs/ai/running-bigdl-deep-learning-for-apache-spark-on-aws/

BigDL on Alibaba* Cloud
E-MapReduce*
https://yq.aliyun.com/articles/73347

BigDL on CDH* and Cloudera* Data Science Workbench*
http://blog.cloudera.com/blog/2017/04/bigdl-on-cdh-and-cloudera-data-science-workbench/

BigDL Spark deep learning library VM now available on Microsoft* Azure* Marketplace https://azure.microsoft.com/en-us/blog/bigdl-spark-deep-learning-library-vm-now-available-on-microsoft-azure-marketplace/

BigDL in KMR* Service of Kingsoft* Cloud
https://docs.ksyun.com/read/latest/33/_book/bigDL.html

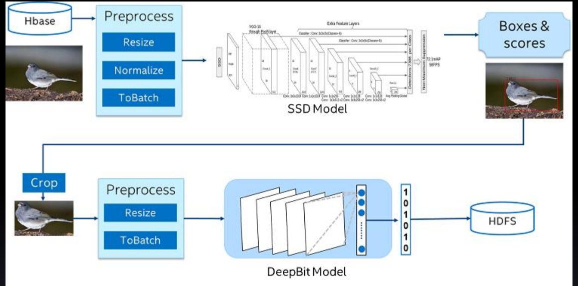
Using BigDL in IBM* Data Science Experience
https://medium.com/ibm-data-science-experience/using-bigdl-in-data-science-experience-for-deep-learning-on-spark-f1cf30ad6ca0

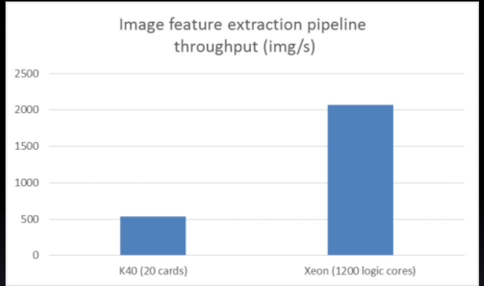
Using BigDL for deep learning with Apache Spark and Google* Cloud Dataproc*
https://cloud.google.com/blog/big-data/2018/04/using-bigdl-for-deep-learning-with-apache-spark-and-google-cloud-dataproc

Intel's BigDL on Databricks*
https://databricks.com/blog/2017/02/09/intels-bigdl-databricks.html

BigDL Shipped in Cray* Urika-XC* Analytics Software Suite
https://www.cray.com/blog/scalable-deep-learning-bigdl-urika-xc-software-suite/

Object Detection and Image Feature Extraction in JD



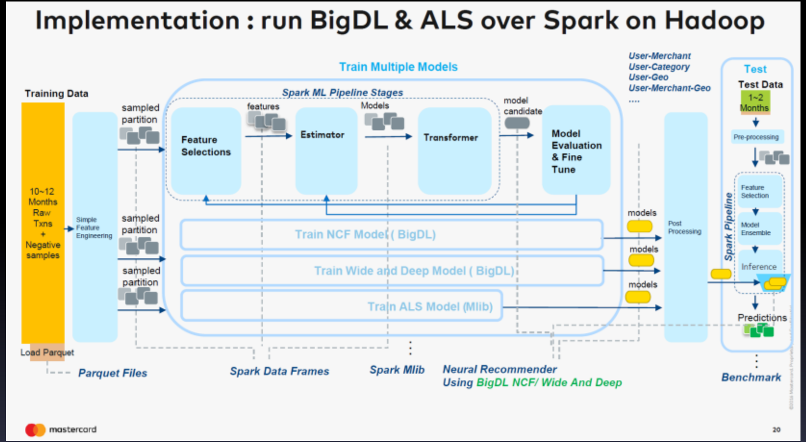


- Reuse existing Hadoop/Spark clusters for deep learning with no changes (image search, IP protection, etc.)
- Efficiently scale out on Spark with superior performance (***3.83x*** speed-up vs. GPU severs) as benchmarked by JD

http://mp.weixin.qq.com/s/xUCkzbHK4K06-v5qUsaNQQ
https://software.intel.com/en-us/articles/building-large-scale-image-feature-extraction-with-bigdl-at-jdcom

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

User-Merchant Propensity Modeling in MasterCard

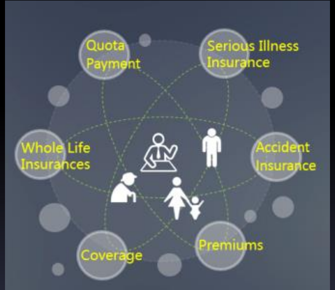


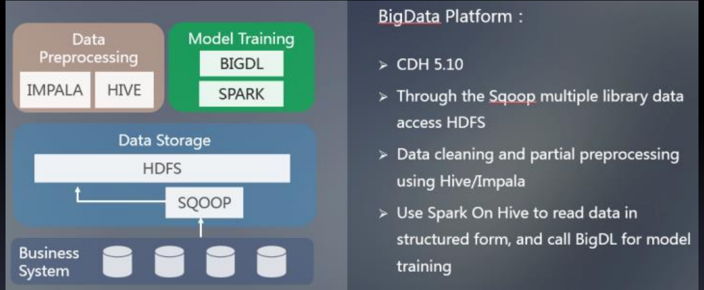
Implementation : run BigDL & ALS over Spark on Hadoop

https://www.ai-expo.net/northamerica/talk/using-deep-learning-intel-bigdl-optimized-personalized-card-linked-offer/
https://conferences.oreilly.com/strata/strata-ca/public/schedule/detail/63897

Neural Recommendation Engine in China Life

Realize re-discovery of life insurance business, accurately and effectively recommend products.



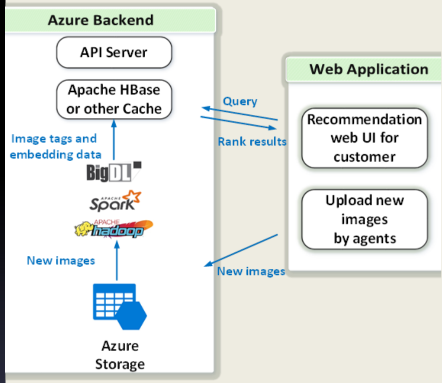


BigData Platform：

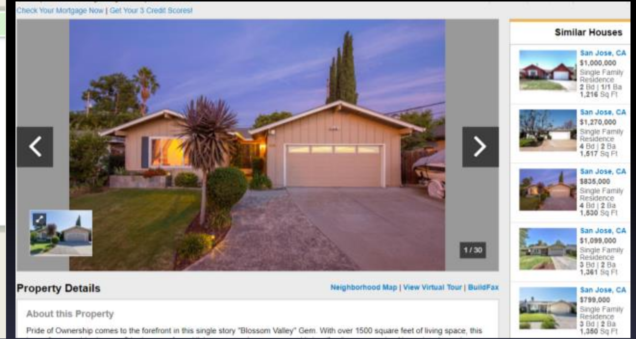
- CDH 5.10
- Through the Sqoop multiple library data access HDFS
- Data cleaning and partial preprocessing using Hive/Impala
- Use Spark On Hive to read data in structured form, and call BigDL for model training

https://strata.oreilly.com.cn/strata-cn/public/schedule/detail/59722?locale=en

Image Similarity Search for MLSListings


Azure Backend
API Server
Apache HBase or other Cache
Image tags and embedding data
BigDL
Spark
hadoop
New images
Azure Storage
Query
Rank results
New images
Web Application
Recommendation web UI for customer
Upload new images by agents

MLSlistings built image-similarity based house recommendations using BigDL on Microsoft Azure



Check Your Mortgage Now | Get Your 3 Credit Scores!

Similar Houses

- San Jose, CA $1,000,000 Single Family Residence 2 Bd | 1/1 Ba 1,216 Sq Ft
- San Jose, CA $1,270,000 Single Family Residence 4 Bd | 2 Ba 1,517 Sq Ft
- San Jose, CA $835,000 Single Family Residence 4 Bd | 2 Ba 1,530 Sq Ft
- San Jose, CA $1,099,000 Single Family Residence 3 Bd | 2 Ba 1,361 Sq Ft
- San Jose, CA $799,000 Single Family Residence 3 Bd | 2 Ba 1,350 Sq Ft

Property Details

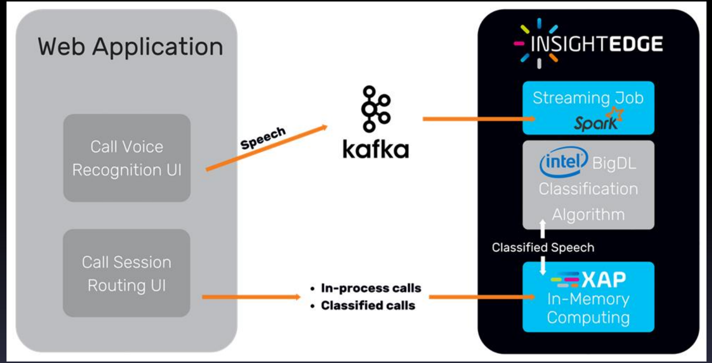
Neighborhood Map | View Virtual Tour | BuildFax

About this Property

Pride of Ownership comes to the forefront in this single story "Blossom Valley" Gem. With over 1500 square feet of living space, this

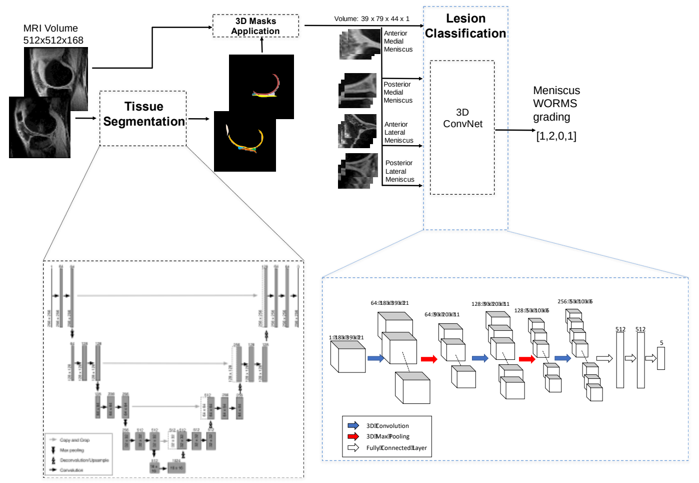
https://software.intel.com/en-us/articles/using-bigdl-to-build-image-similarity-based-house-recommendations

NLP Based Call Center Routing in GigaSpaces



https://blog.gigaspaces.com/gigaspaces-to-demo-with-intel-at-strata-data-conference-and-microsoft-ignite/

3D Medical Image Analysis in UCSF



https://conferences.oreilly.com/strata/strata-ca/public/schedule/detail/64023

Partner With Us

- Use Analytics-Zoo & Share your Experience
- Use Intel Optimized Libraries & Frameworks
- Leverage Intel Developer Zone Resources

Source code: https://github.com/intel-analytics/analytics-zoo/
Documents: https://analytics-zoo.github.io/

intel®

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.