Accelerate Innovation in the Enterprise with Distributed ML / DL

Nanda Vijaydev – BlueData (recently acquired by HPE)
The AI Conference, New York
• AI, Machine Learning (ML), and Deep Learning (DL)
• Example Enterprise Use Cases
• Deployment Challenges for Distributed ML / DL
• Distributed TensorFlow and Horovod on Containers with Intel Xeon processors and Intel MKL
• Lessons Learned and Key Takeaways
AI, Machine Learning, and Deep Learning
Let’s Get Grounded…What is AI?
What are Machine Learning and Deep Learning?

Artificial intelligence (AI)
Mimics human behavior. Any technique that enables machines to solve a task in a way like humans do.

Example: Siri

Deep learning (DL)
Subset of ML, using deep artificial neural networks as models, inspired by the structure and function of the human brain.

Example: Self-driving car

Machine learning (ML)
Algorithms that allow computers to learn from examples without being explicitly programmed.

Example: Google Maps
Why Should You Be Interested in AI / ML / DL?

Everyone wants AI / ML / DL and advanced analytics....

AI and advanced analytics represent 2 of top 3 CIO priorities

AI and advanced analytics infrastructure could constitute 15-20% of the market by 2021\(^1\)

Enterprise AI adoption 2.7X growth in last 4 years\(^2\)

....but face many challenges

Use cases
New roles, skill gaps
Culture and change
Data preparation
Legacy infrastructure

\(^1\) IDC, Goldman Sachs, HPE Corporate Strategy, 2018
\(^2\) Gartner - “2019 CIO Survey: CIOs Have Awoken to the Importance of AI”
### Key Questions Remain ...

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What opportunities does AI bring to your business? What are the major use cases?</td>
<td></td>
</tr>
<tr>
<td>How do you get started with gaining intelligence with your data?</td>
<td></td>
</tr>
<tr>
<td>What is the best way to prepare your company for a data-centric and AI future?</td>
<td></td>
</tr>
<tr>
<td>How do you integrate your AI and data ecosystem for ML / DL and advanced analytics?</td>
<td></td>
</tr>
<tr>
<td>How do you modernize, consume, and prepare your EDW or Hadoop big data foundation for AI?</td>
<td></td>
</tr>
</tbody>
</table>
AI / ML / DL Adoption in the Enterprise

**Financial services**
- Fraud detection, ID verification

**Government**
- Cyber-security, smart cities and utilities

**Energy**
- Seismic and reservoir modeling

**Retail**
- Video surveillance, shopping patterns

**Health**
- Personalized medicine, image analytics

**Consumer tech**
- Chatbots

**Service providers**
- Media delivery

**Manufacturing**
- Predictive and prescriptive maintenance
Example Enterprise Use Cases
### Financial Services Use Cases

#### Wide Range of ML / DL Use Cases for Wholesale / Commercial Banking, Credit Card / Payments, Retail Banking, etc.

<table>
<thead>
<tr>
<th>Fraud Detection</th>
<th>Risk Modeling &amp; Credit Worthiness Check</th>
<th>CLV Prediction and Recommendation</th>
<th>Customer Segmentation</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-Time Transactions</td>
<td>Loan Defaults</td>
<td>Historical Purchase View</td>
<td>Behavioral Analysis</td>
<td>Image Recognition</td>
</tr>
<tr>
<td>Credit Card</td>
<td>Delayed Payments</td>
<td>Pattern Recognition</td>
<td>Understanding</td>
<td>NLP</td>
</tr>
<tr>
<td>Merchant</td>
<td>Liquidity</td>
<td>Retention Strategy</td>
<td>Customer Quadrant</td>
<td>Security</td>
</tr>
<tr>
<td>Collusion</td>
<td>Market &amp; Currencies</td>
<td></td>
<td>Effective Messaging &amp; Improved Engagement</td>
<td>Video Analysis</td>
</tr>
<tr>
<td>Impersonation</td>
<td>Purchases and Payments</td>
<td></td>
<td>Targeted Customer</td>
<td></td>
</tr>
<tr>
<td>Social Engineering Fraud</td>
<td>Time Series</td>
<td></td>
<td>Support</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Enhanced Retention</td>
<td></td>
</tr>
</tbody>
</table>

**CLV**: Customer Lifetime Value
Fraud Detection Use Case

• One of the most common use cases for ML / DL in Financial Services is to detect and prevent fraud

• This requires:
  – Distributed Big Data processing frameworks such as Spark
  – ML / DL tools such as TensorFlow, H2O, and others
  – Continuous model training and deployment
  – Multiple large data sets
Fraud Detection Use Case (cont’d)

• Data science teams need the ability to create distributed ML / DL environments for sandbox as well as trial and error experimentation

• This requires:
  – Hardware acceleration (e.g. Xeon, MKL)
  – Multiple different ML / DL and data science tools
  – Fast and repeatable deployment of clusters
ML / DL in Healthcare – Use Cases

• Precision Medicine and Personal Sensing
  – Disease prediction, diagnosis, and detection (e.g. genomics research)
  – Using data from local sensors (e.g. mobile phones) to identify human behavior

• Electronic Health Record (EHR) correlation
  – “Smart” health records

• Improved Clinical Workflow
  – Decision support for clinicians

• Claims Management and Fraud Detection
  – Identify fraudulent claims

• Drug Discovery and Development
Use Case: Precision Medicine

• Many types of data
  – Genomic
  – Microbiome
  – Epigenome
  – Etc.

• Huge volumes of data (petabytes > exabytes)
PERSONALIZED MEDICINE COMES TO CYSTIC FIBROSIS

Ivacaftor targets G551D mutation with success

A phase III study of an experimental, targeted cystic fibrosis (CF) treatment shows sustained lung function improvement for patients with CF with a certain mutation, according to a report published in November in the New England Journal of Medicine [Ramsay et al., 2011].

The experimental drug ivacaftor, also known as VX-770, is the first personalized treatment for CF. It targets an underlying genetic cause of CF, the G551D-CFTR mutation, not just its symptoms. About 4-5% of patients with CF have at least one such mutation.

Ivacaftor produced substantial lung function improvement in patients with CF age 12 and older after two weeks. It sustained this improvement through the 48 weeks of the study. Ivacaftor produced a 10.6 percentage point improvement in forced expiratory volume in one second (FEV1) over placebo at 24 weeks, reported by Bonnie Ramsay, MD, Director of the Center...
Deployment Challenges for Distributed ML / DL
Why Distributed ML / DL?

- Large Data Volumes
- Speed
- Fault Tolerance
Distributed ML / DL – Challenges

• Complexity, lack of repeatability and reproducibility across environments
• Sharing data, not duplicating data
• Need agility to scale up and down compute resources
• Deploying multiple distributed platforms, libraries, applications, and versions
• One size environment fits none
• Need a flexible and future-proof solution
Example Deployment Challenges

• How to run clusters on heterogeneous host hardware
  – CPUs and GPUs, including multiple GPU versions
• How to maximize use of expensive hardware resources
• How to minimize manual operations
  – Automating the cluster creation and deployment process
  – Creating reproducible clusters and reproducible results
  – Enabling on-demand provisioning and elasticity
Example Deployment Challenges

• How to support the latest versions of software
  – Deployment complexity and upgrades
  – Version compatibility

• How to ensure enterprise-class security
  – Network, storage, user authentication, and access
Docker is software that performs operating-system-level virtualization also known as containerization.

Containerization allows the existence of multiple instances on a server.

Source: https://en.wikipedia.org/wiki/docker_(software)
Distributed ML / DL and Containers

- ML / DL applications are compute hardware intensive
- They can benefit from the flexibility, agility, and resource sharing attributes of containerization
- But care must be taken in how this is done, especially in a large-scale distributed environment
AI-Driven Solutions for the Enterprise

Example Industry Use Cases

<table>
<thead>
<tr>
<th>Solutions</th>
<th>Example Industry Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science</td>
<td>Fraud Detection</td>
</tr>
<tr>
<td>and ML / DL Tools</td>
<td>Genome Research</td>
</tr>
<tr>
<td>Data Platforms</td>
<td>Customer 360</td>
</tr>
<tr>
<td>Data</td>
<td>Video Surveillance</td>
</tr>
<tr>
<td>IT</td>
<td></td>
</tr>
</tbody>
</table>

- **HDFS/NFS**
  - Data Store

- **Data Duplication**

- **Cloud**

- **User Access**

- **Security**

- **Time to Deploy**

- **Multi-Tenant**
Turnkey Container-Based Solution

Data Scientists  Developers  Data Engineers  Data Analysts

BlueData EPIC™ Software Platform

ElasticPlane™ – Self-service, multi-tenant clusters
IOBoost™ – Extreme performance and scalability
DataTap™ – In-place access to data on-prem or in the cloud

Compute
CPUs  GPUs
Storage
NFS  HDFS

On-Premises  Public Cloud

Big Data Tools
Spark  Kafka  Cloudera

ML/DL Tools
H2O AI  BigDL  TensorFlow

Data Science Tools
jupyter  R

BI/Analytics Tools

Bring-Your-Own

Compute

Storage

On-Premises
Public Cloud
TensorFlow and Horovod on Containers with Intel Xeon processors and Intel MKL
Distributed TensorFlow – Concepts

• Running TensorFlow training in parallel, on multiple devices
• Goal is to improve accuracy and speed
• Different layers may be trained on different nodes (model parallelism)
• Same model can be applied on different subsets of data, in different nodes (data parallelism)

https://towardsdatascience.com/distributed-tensorflow-using-horovod-6d572f8790c4
Distributed TensorFlow – Schemes

- Data parallelism implementation
  - Needs to sync model parameters
  - Uses a centralized or decentralized scheme to communicate parameter update

- Centralized schemes use Parameter Server to communicate updates to parameters (gradients) between nodes

- Decentralized schedules use ring-allreduce scheme

- Horovod is an open source framework developed by Uber that supports allreduce
Meet Horovod

- Distributed training framework for
  - Tensorflow, PyTorch, Keras
- Separates infrastructure capabilities from ML
- Installs easily on existing ML framework
  - pip install horovod
- Uses bandwidth optimal communication protocol
  - RDMA, InfiniBand if available

horovod.ai
 TensorFlow with Horovod on Docker

Docker Containers

Horovod cluster on multiple containers, and machines

- MPI 3.1.3
- TensorFlow 1.7*
- MKL

Xeon®

Shared Data
TensorFlow with Horovod

- `tensorflow_wrd2vec.py` from git [https://github.com/horovod/horovod](https://github.com/horovod/horovod) examples
- Data comes from shared NFS mounts, automatically surfaced by BlueData into containers
- Passwordless ssh setup during cluster creation
- All prerequisites installed on all nodes, including:
  - MKL – Math Kernel Library
  - tensorflow, pytorch, scikit-learn, ... (compute frameworks)
  - openmpi (To distribute the job)
  - tensorboard for visualization
### App Store with Pre-Built ML / DL Images

<table>
<thead>
<tr>
<th><strong>TensorFlow</strong></th>
<th><strong>TensorFlow 1.9 with CUDA</strong></th>
<th><strong>Installed</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TensorFlow</strong></td>
<td><strong>TensorFlow Serving 1.12</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>jupyterhub</strong></td>
<td><strong>JupyterHub 0.9.4 with Sparkmagic</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>TensorFlow with Horovod 0.16.0</strong></td>
<td><strong>Installed</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Spark</strong></td>
<td><strong>Spark 2.4</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>H2O.ai</strong></td>
<td><strong>H2O 3.22.18</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>Sparkling Water</strong></td>
<td><strong>Sparkling Water with Spark 2.4</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>H2O Driverless AI 1.5.4</strong></td>
<td><strong>Installed</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Caffe2</strong></td>
<td><strong>Caffe 2.0.8.1 with Anaconda</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>Keras</strong></td>
<td><strong>Keras 2.2.1</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>mxnet</strong></td>
<td><strong>MXNet 1.4.0</strong></td>
<td><strong>Installed</strong></td>
</tr>
<tr>
<td><strong>PyTorch</strong></td>
<td><strong>PyTorch 1.0.1 with CUDA</strong></td>
<td><strong>Installed</strong></td>
</tr>
</tbody>
</table>

**Docker images for multiple applications and versions**

**Ability to create and add new images**
mpirun \(-np 4\) \\
\hspace{1em} \text{--allow-run-as-root} /
\hspace{1em} -d \hspace{1em} -H \text{bluedata-302.bdlocal:2,bluedata-301.bdlocal:4} / 
\hspace{1em} -bind-to none -map-by slot / 
\hspace{1em} -x LD_LIBRARY_PATH / 
\hspace{1em} -x PATH / 
\hspace{1em} -mca pml ob1 / 
\hspace{1em} -mca btl ^openib python tensorflow_word2vec_logs.py
Lessons Learned and Key Takeaways
Lessons Learned and Takeaways

• Enterprises are using ML / DL today to solve difficult problems (example use cases: fraud detection, disease prediction)

• Distributed ML / DL in the enterprise requires a complex stack, with multiple different tools (TensorFlow is one popular option)

• The only constant is change … be prepared
  – Business needs, use cases, and tools will constantly evolve

• Deployments are challenging, with many potential pitfalls
  – Containerization can deliver agility and cost saving benefits
Lessons Learned and Takeaways

• Leverage a flexible, scalable, and elastic platform for success
  – BlueData provides a turnkey container-based platform for large-scale distributed AI / ML / DL in the enterprise
  – Enterprise-grade security and performance, proven in production at leading Global 2000 organizations
  – Decouple compute from storage for greater efficiency, and deploy on-premises, in a hybrid model, or multi-cloud
  – Save time, save money, and accelerate innovation
Thank You

To learn more, visit the **BlueData** booth in the Expo Hall

[www.bluedata.com](http://www.bluedata.com)