CaffeOnSpark: Deep Learning on Spark Cluster

Andy Feng, Jun Shi and Mridul Jain
Yahoo! Inc.

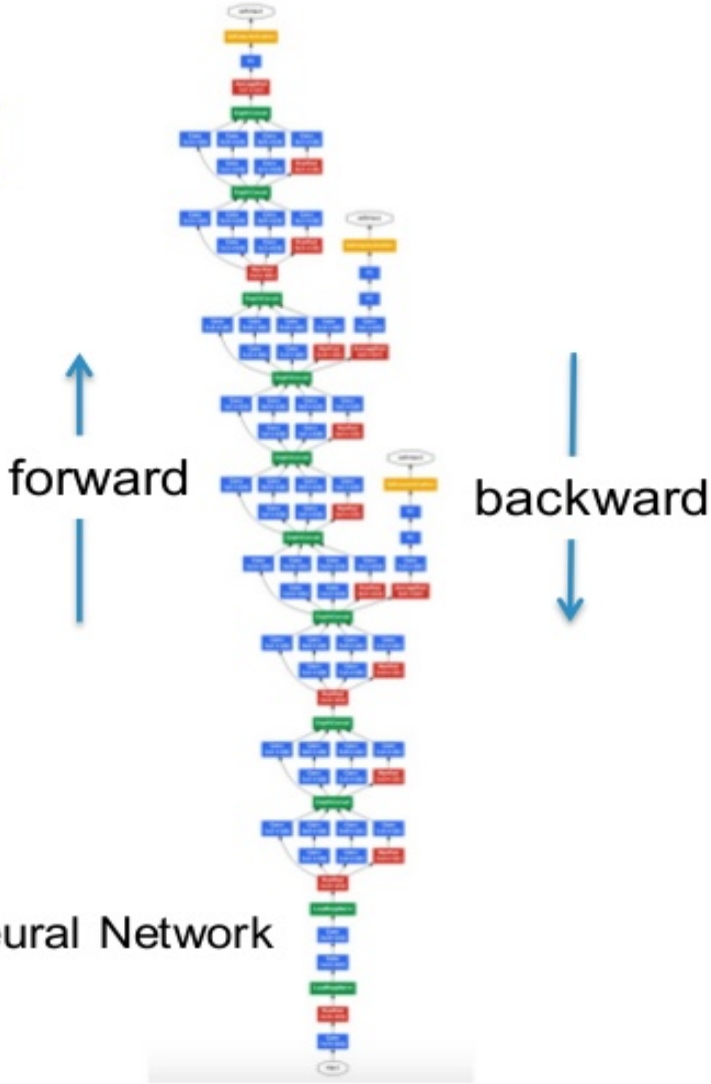
SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO

Agenda

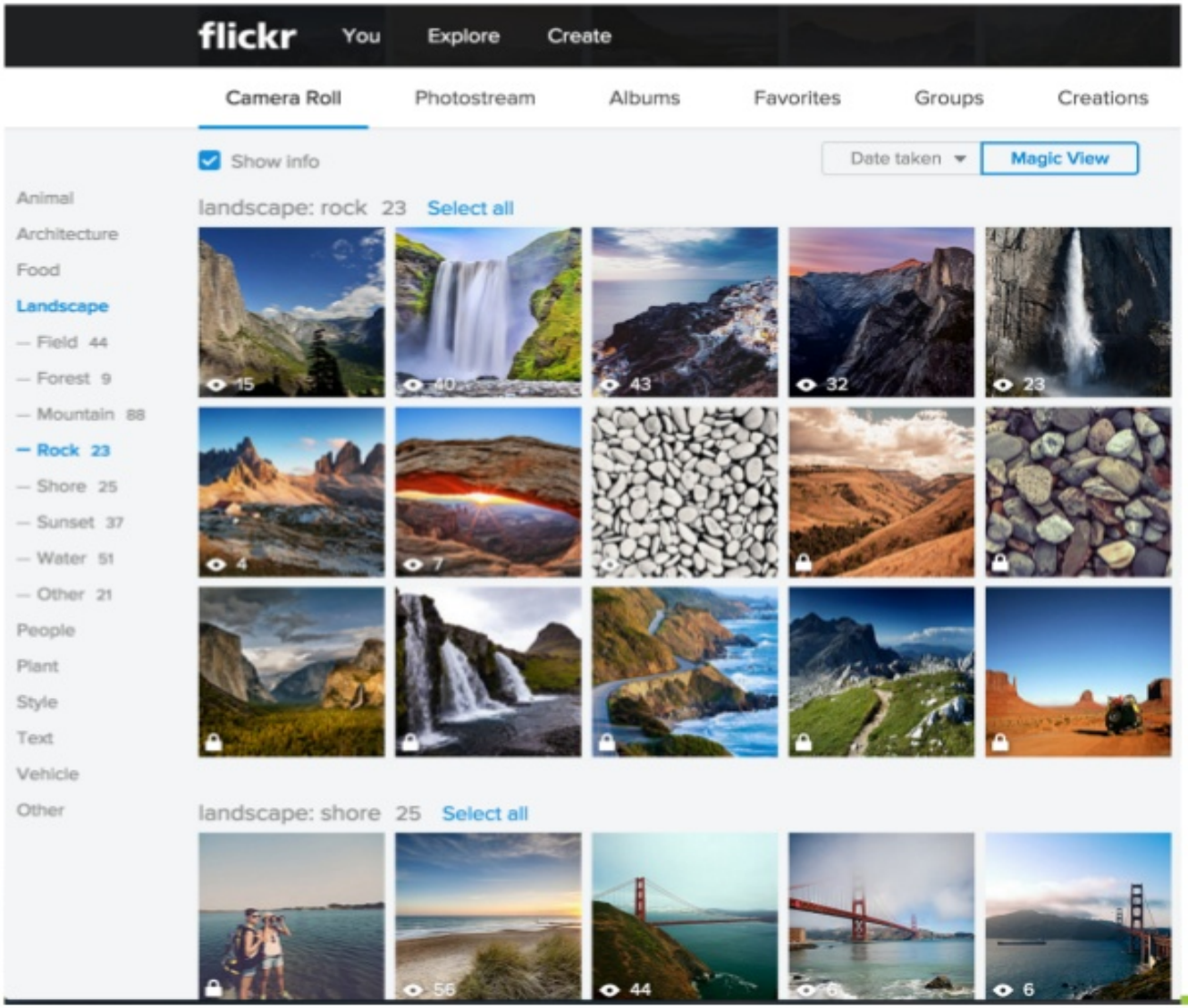
- Why Deep Learning on Spark?
- CaffeOnSpark
  - Architecture
  - API: Scala + Python
- Demo
  - CaffeOnSpark on Python Notebook

Deep Learning

Handwritten digits (MNIST)



Deep Neural Network

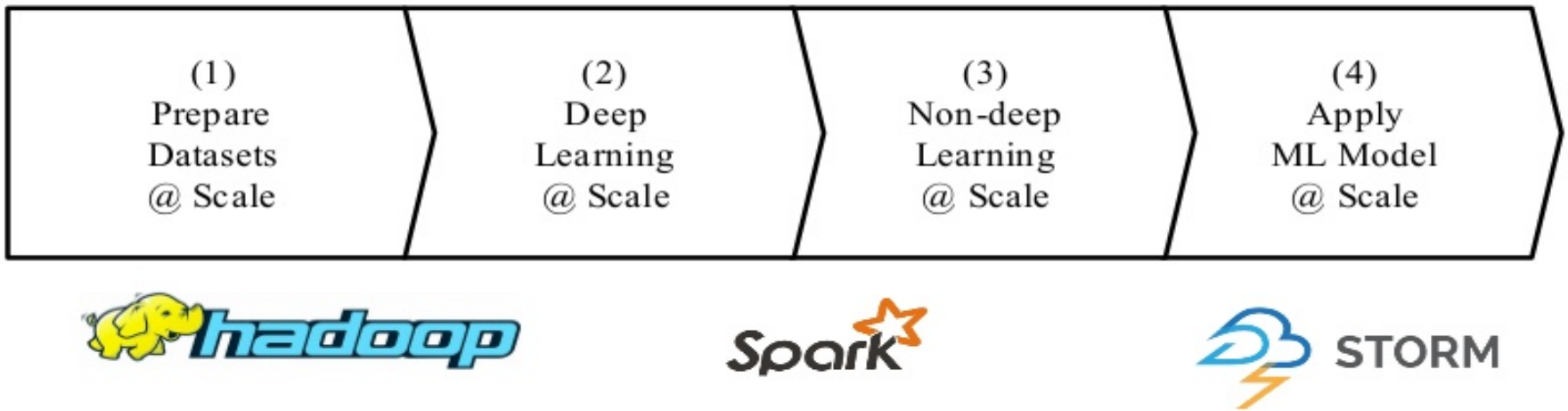


Flickr Magic View:

https://flickr.com/cameraroll

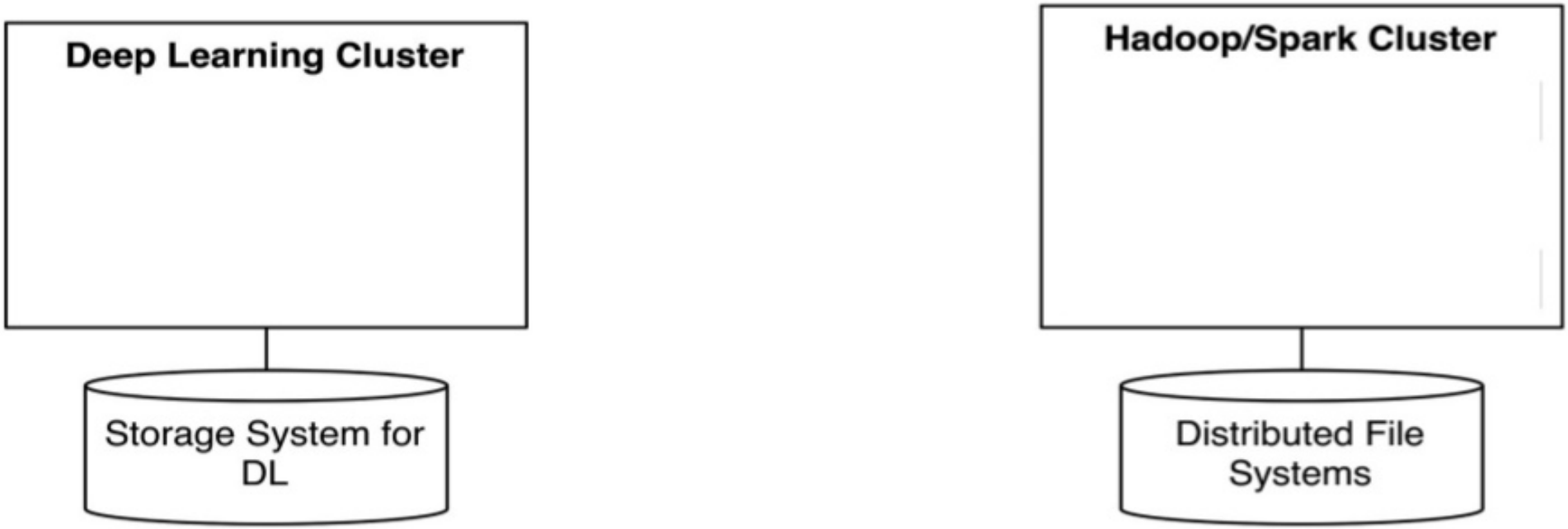
- Photos organized according to 70 categories
- Empowered by deep learning & machine learning

Flickr DL/ML Pipeline



* http://bit.ly/1KIDfof by Pierre Garrigues, Deep Learning Summit 2015

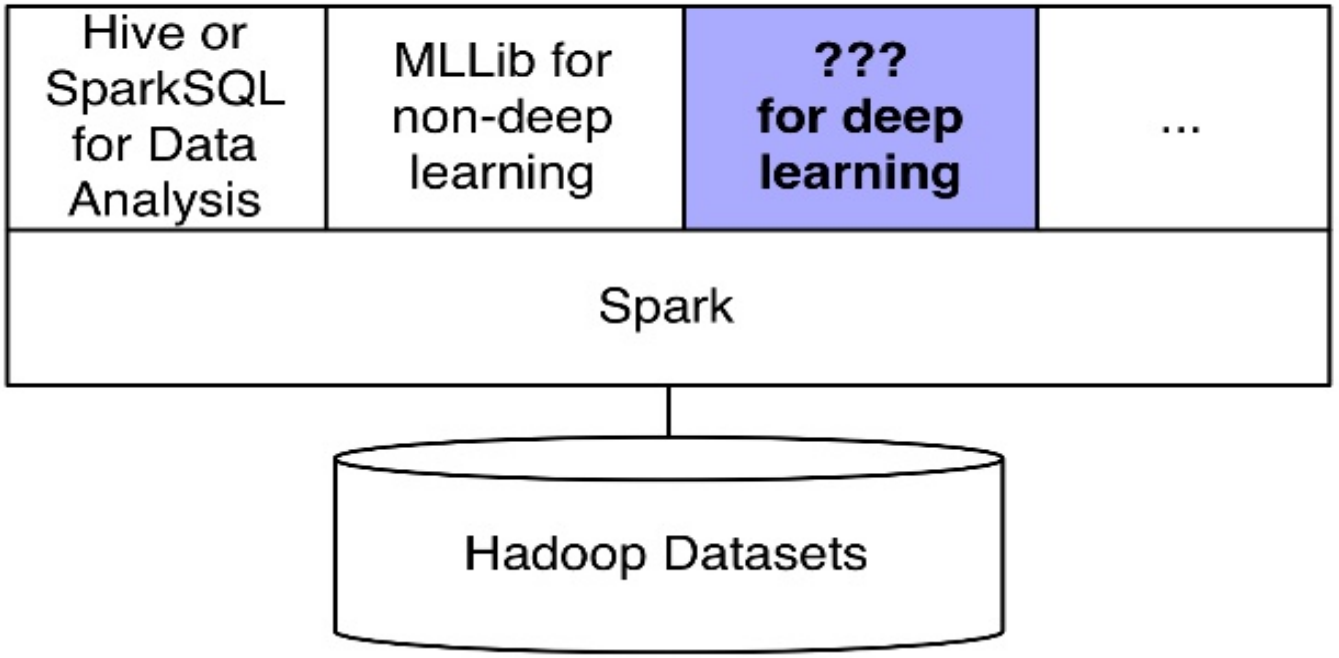
Deep Learning vs. Spark



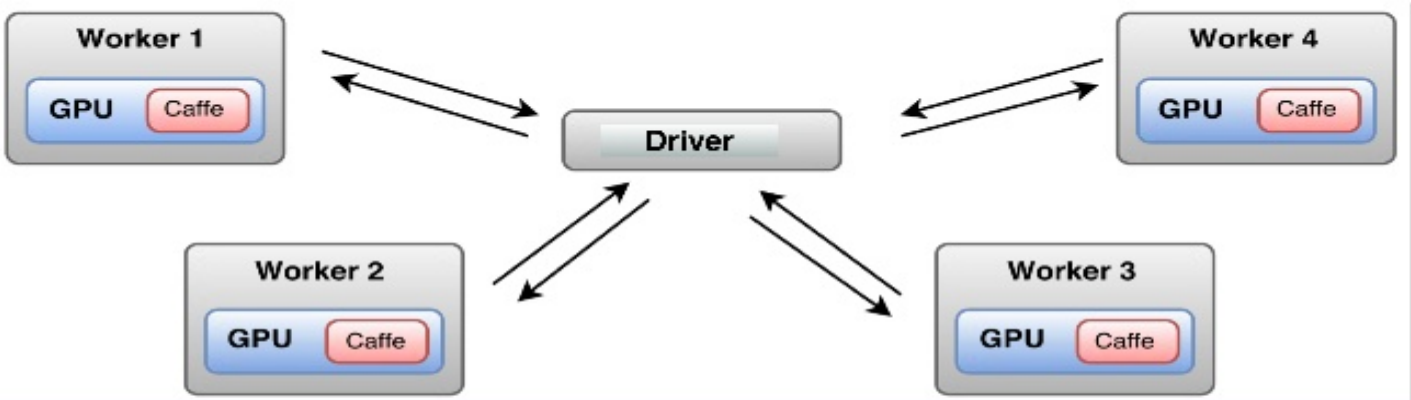
Deep Learning Frameworks

- Theano
- Torch
- Caffe
  - Popular choice for vision community
  - Widely used in Yahoo
- TensorFlow
- …

Deep Learning on Spark



Related Work: SparkNet & DL4J



REPEAT

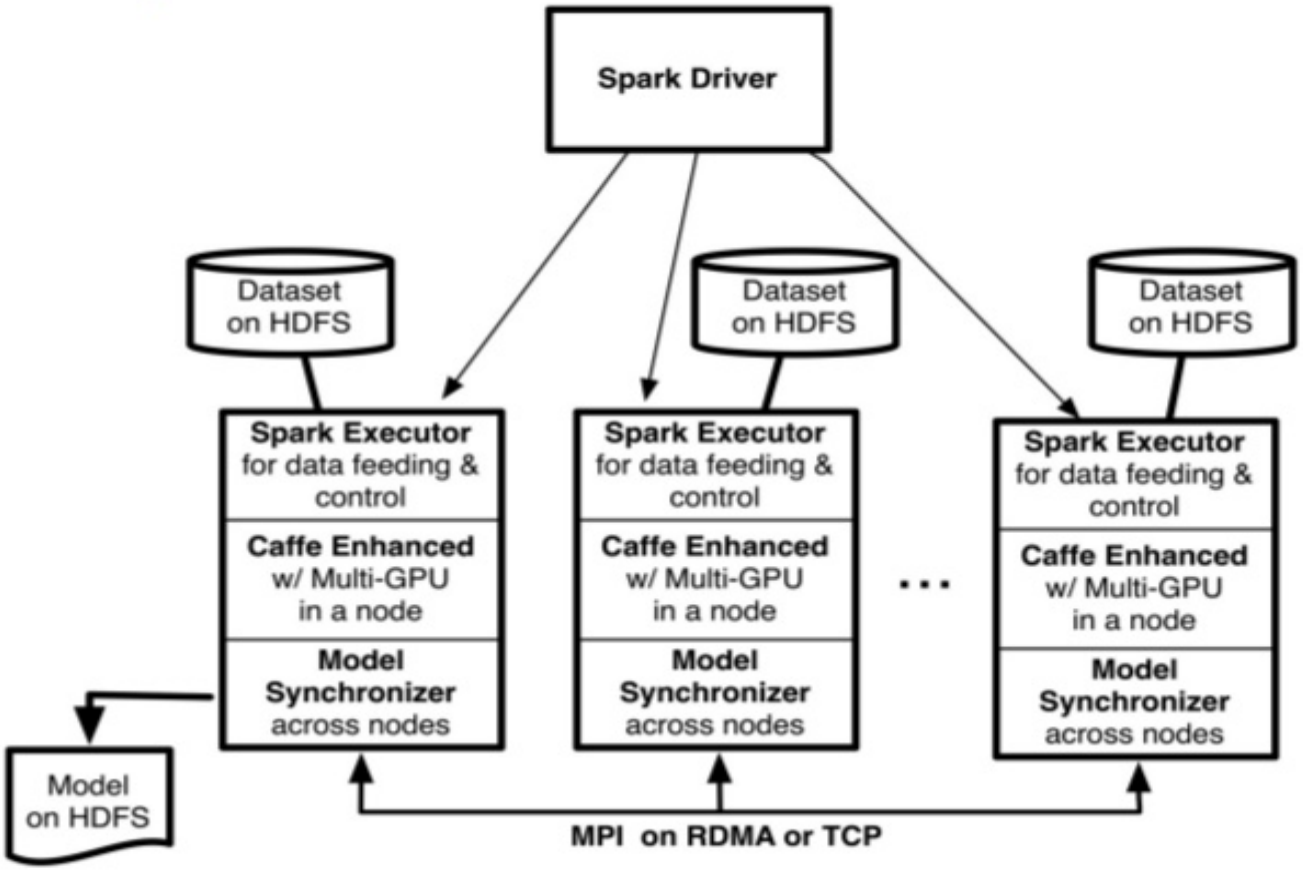
1) [driver] sc.broadcast(model) to executors
2) [executor] apply DL training against a mini-batch of dataset to update models locally
3) [driver] aggregate(models) to produce a new model

CaffeOnSpark Open Sourced

github.com/yahoo/CaffeOnSpark

- Apache 2.0 license
- Distributed deep learning
  - GPU or CPU
  - Ethernet or InfiniBand
- Easily deployed on public cloud or private cloud

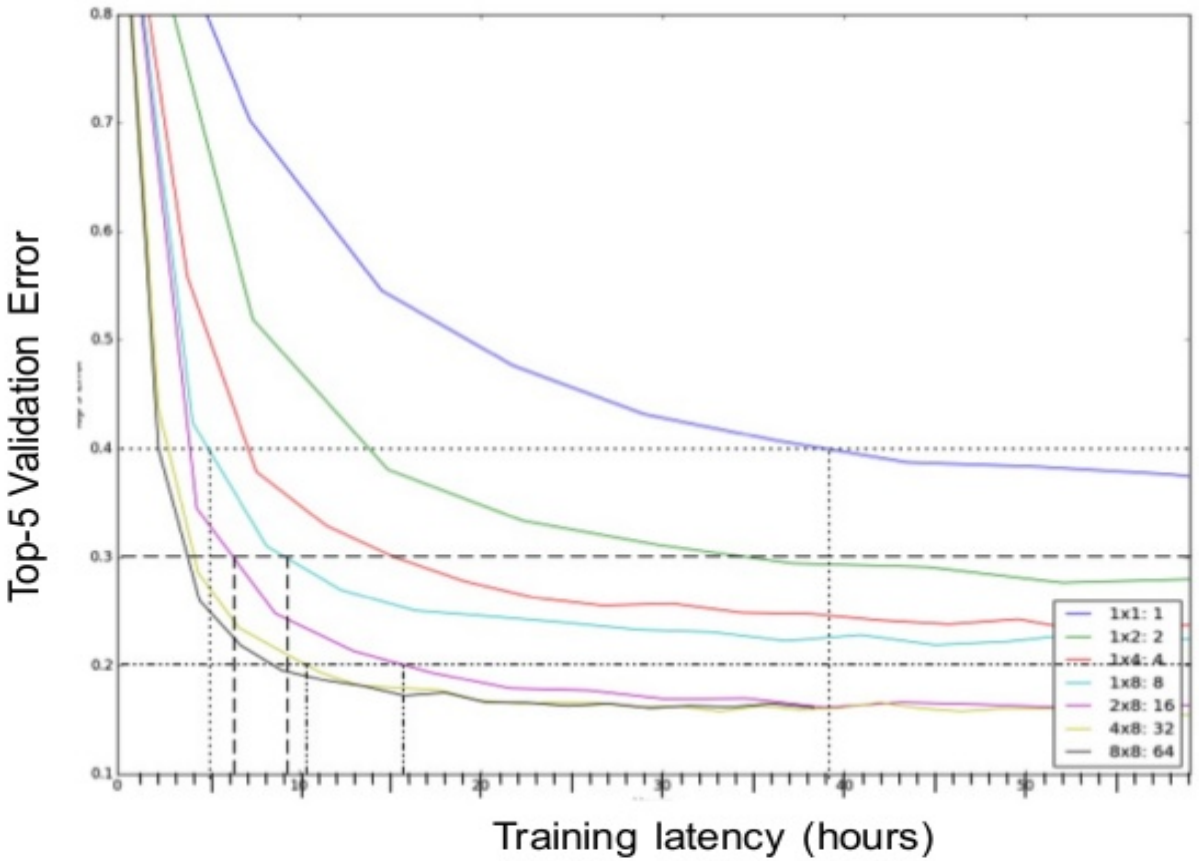
CaffeOnSpark: Scalable Architecture



CaffeOnSpark: Deployment Options

- Single node
  - Spark-submit –master local
- Multiple nodes w/ ethernet connection
  - Spark-submit –master URL –connection ethernet
  - Ex. EC2
- Multiple nodes w/ Infiniband connection
  - Spark-submit –master URL –connection infiniband
  - Ex., Yahoo Hadoop cluster

Deep Learning: 19x Speedup (est.)



CaffeOnSpark: DL Made Easy

Spark CLI

- spark-submit
  --num-executors #_*Processes*
  --class com.yahoo.ml.CaffeOnSpark
  caffe-on-spark.jar
  -devices #_*gpus_per_proc*
  -conf *solver_config_file*
  -model model_file
  -train | -test | -feature

Caffe Configuration

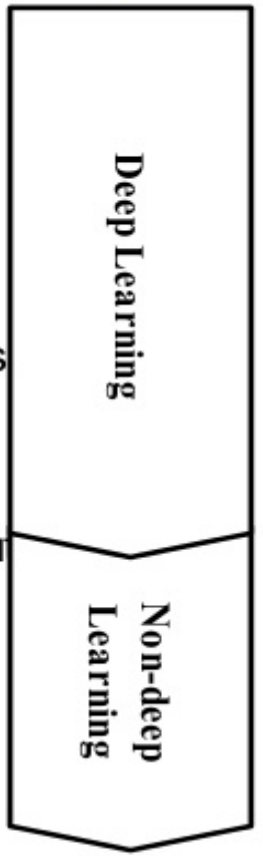
```
layer {
  name: "data"
  type: "MemoryData"
  source_class="com.yahoo.ml.caffe.LMDB"
  memory_data_param {
    source: "hdfs:///mnist/trainingdata/"
    batch_size: 64;
    channels: 1;
    height: 28;
    width: 28;
  }
  …
}
```

CaffeOnSpark: One Program (Scala)

http://bit.ly/21ZY1c2

```
cos = new CaffeOnSpark(ctx)  conf = new Config(ctx, args).init()

// (1) training DL model
dl_train_source = DataSource.getSource(conf, true)  cos.train(dl_train_source)
// (2) extract features via DL
lr_raw_source = DataSource.getSource(conf, false)  ext_df = cos.features(lr_raw_s
// (3) apply ML
lr_input=ext_df.withColumn("L", cos.floats2doubleUDF(ext_df(conf.label)))
.withColumn("F", cos.floats2doublesUDF(ext_df(conf.features(0))))  lr = new
LogisticRegression().setLabelCol("L").setFeaturesCol("F")  lr_model = lr.fit(lr_inpu
```



CaffeOnSpark: One Notebook (Python)

http://bit.ly/1REZ0cN

Feature Extraction

In [50]:
```
args['features']='accuracy,ip1,ip2'
args['label']='label'
cfg=Config(sc,args)
```

In [52]:
```
dl_feature_source = DataSource(sc).getSource(cfg,False)
```

In [54]:
```
f=cos.features(dl_feature_source)
```

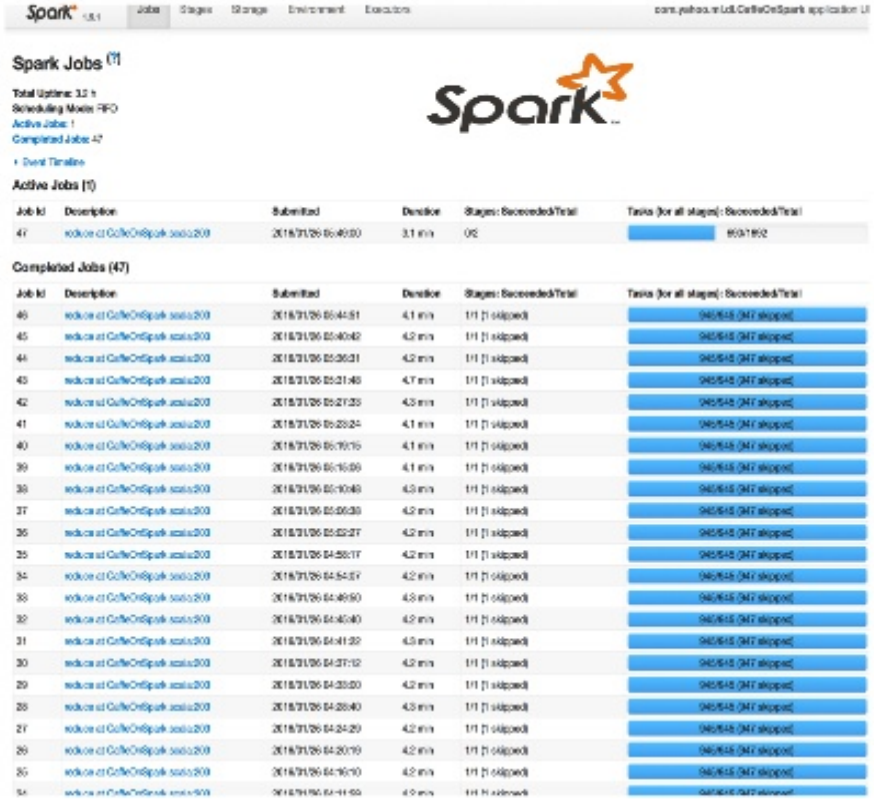
In [55]:
```
f.show(5)
```

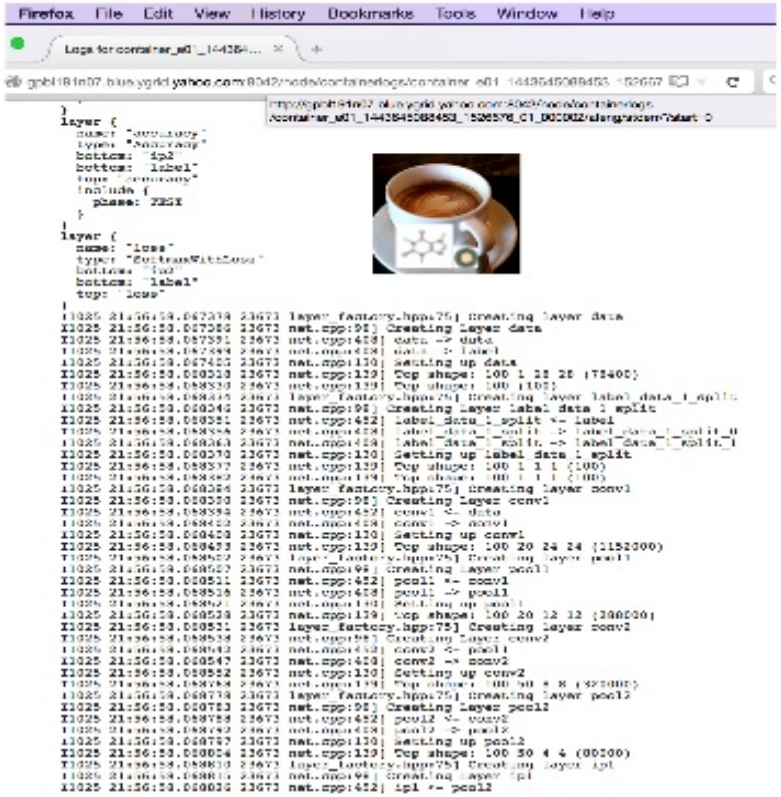
```
+--------+--------+--------------------+--------------------+-----+
|SampleID|accuracy|                 ip1|                 ip2|label|
+--------+--------+--------------------+--------------------+-----+
|00000000|   [1.0]|[-0.0, 3.109636, ...|[-0.6478175, -1.4...|[7.0]|
|00000001|   [1.0]|[1.3683326, -0.0,...|[2.0906663, 1.048...|[2.0]|
|00000002|   [1.0]|[1.5641443, -0.0,...|[-0.773368, 10.61...|[1.0]|
|00000003|   [1.0]|[-0.0, 1.9505613,...|[16.46351, -6.917...|[0.0]|
|00000004|   [1.0]|[0.5979191, 0.075...|[-0.48371825, -2....|[4.0]|
+--------+--------+--------------------+--------------------+-----+
```

In [45]:
```
dl_train_source = DataSource(sc).getSource(cfg,True)
```

In [46]:
```
cos.train(dl_train_source)
```

CaffeOnSpark: UI & Logs





Demo: CaffeOnSpark on EC2

- https://github.com/yahoo/CaffeOnSpark/wiki
  - Get started on EC2
  - Python for CaffeOnSpark

Summary

- CaffeOnSpark open sourced
  - https://github.com/yahoo/CaffeOnSpark
  - Empower Flickr and other Yahoo services
  - Scalable DL made easy

THANK YOU.

bigdata@yahoo-inc.com

SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE
JUNE 6-8, 2016 SAN FRANCISCO