MLflow and MLeap

Andre Mesarovic
Sr. Resident Solutions Architect at Databricks
2019-12-03

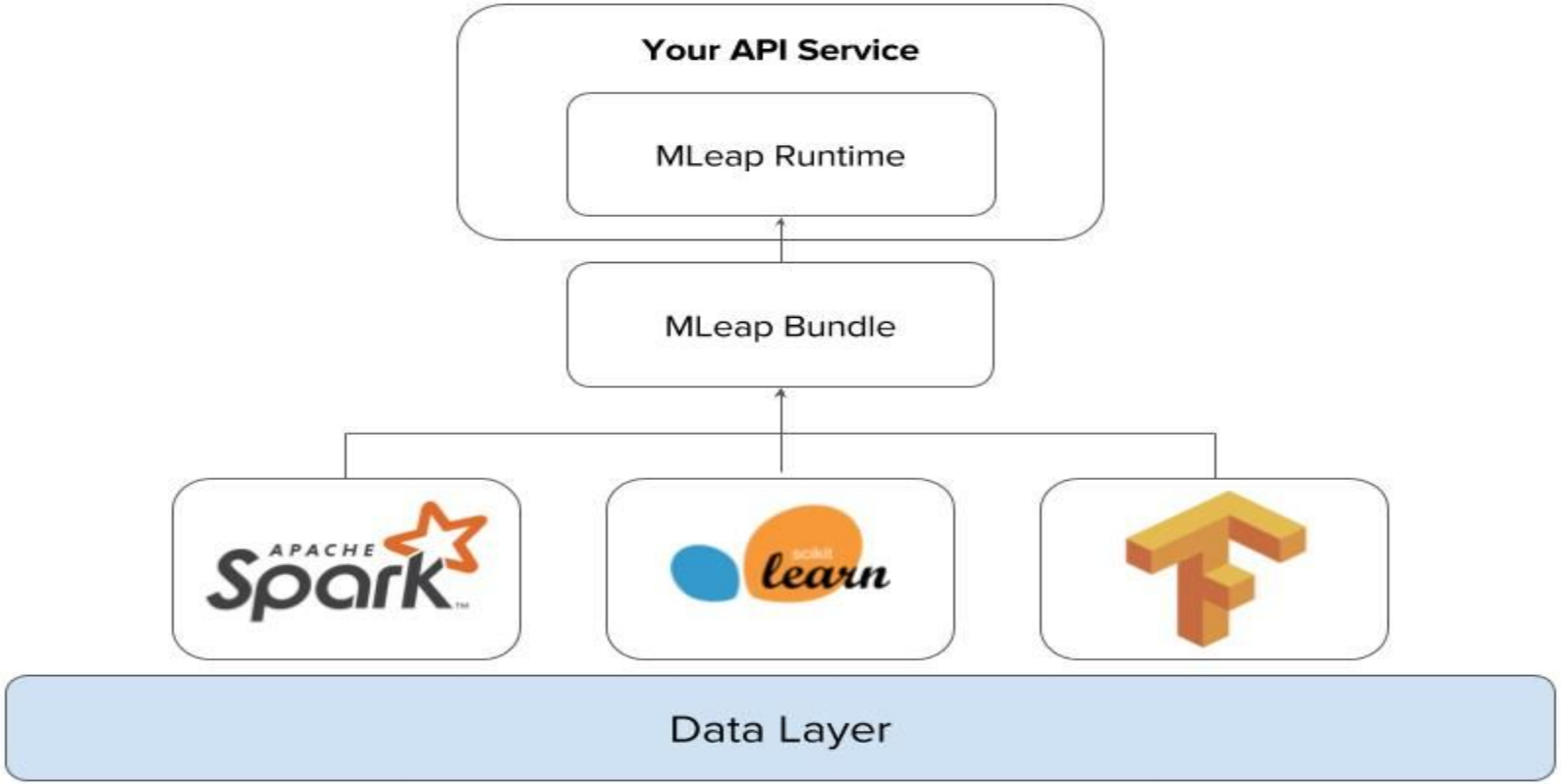
Problem statement

- Saving a model in Spark ML format means you need Spark to deserialize and score it
- Spark is not well suited for real-time scoring
- In production environments Spark might not be available
- MLeap decouples training from serving environment

What is MLeap?

- MLeap is:
  - Common serialization format for models
  - Execution engine for machine learning pipelines
- MLeap supports Spark, Scikit-learn and Tensorflow for training pipelines and exporting them to an MLeap bundle.
- MLeap bundle can be deserialized as:
  - SparkBundle - Use Spark for batch scoring
  - MLeapBundle - Use MLeap runtime to power realtime API services with *no* Spark dependencies

MLeap Overview



MLeap is ambitious

- MLeap claims it implements much of the inference part of Spark, TensorFlow and Scikit-learn
- This is not a trivial undertaking!
- Supported transformers:
  - https://mleap-docs.combust.ml/core-concepts/transformers/support.html
- Parity tests to validate the fidelity of the transformation?
- Is the MLeap variant of the Spark ML model consistent?
- Uber's Michelangelo opted out from using MLeap
- 51,121 lines of Scala code

Bundle Formats

- Two bundle formats: file and jar
  - "file:/tmp/model-bundle"
  - "jar:file:/tmp/model-bundle.zip"
- File format is an exploded directory
- Jar format is a zipped version
- Make sure path exists otherwise you get a nice NPE
- Bundle can be either JSON or Protobuf

Sample Bundle File

root/DecisionTreeRegressor_7f1c4611c78f.node/tree.json


```
{"type":"internal","split":{"type":"continuous","featureIndex":10,"threshold":10.64999}}
{"type":"internal","split":{"type":"continuous","featureIndex":1,"threshold":0.2375}}
{"type":"internal","split":{"type":"continuous","featureIndex":5,"threshold":26.5}}
{"type":"internal","split":{"type":"continuous","featureIndex":0,"threshold":8.45}}
{"type":"internal","split":{"type":"continuous","featureIndex":6,"threshold":166.5}}
{"type":"leaf","values":[5.6982248520710055]}
{"type":"leaf","values":[6.666666666666667]}
{"type":"internal","split":{"type":"continuous","featureIndex":8,"threshold":3.105}}
```

MLeap Scoring Performance

- MLeap is fast
- Both for model loading and model scoring
- MLeap loads models faster than Spark ML
- Scoring can be up to 26x faster
- Links:
  - https://github.com/combust/mleap/tree/master/mleap-benchmark

MLflow MLeap vs Spark Real-time Scoring

- Mean latency for scoring for MLflow SageMaker container:
  - Spark ML: 0.158 seconds
  - MLeap: 0.006 seconds
  - Hence MLeap is 26x faster

MLeap Model Load Performance

Small test on laptop (seconds)

MLflow 1.3.0

```
spark - 9.55, 15.28, 10.36
mleap - 0.35,  1.11,  0.60
```

MLflow 1.4.0

```
spark - 2.92, 2.62, 2.82, 2.72
mleap - 0.27, 0.30, 0.27, 0.30
```

MLeap Examples

Scala

- Write Spark model as bundle
- Read as SparkBundle
- Read as MLeapBundle

Python

- Write Spark model as bundle
- Read as SparkBundle
- ~~Read as MLeapBundle~~ - only available in Scala

MLeap without MLflow Examples

Python - Write Bundle

```
from mleap.pyspark.spark_support import SimpleSparkSerializer

val predictions = model.transform(test_data)
model.serializeToBundle(“jar:file:/tmp/model.zip”, predictions)
```

Python - Read as SparkBundle

```
from pyspark.ml import PipelineModel
from mleap.pyspark.spark_support import SimpleSparkSerializer

model = PipelineModel.deserializeFromBundle(“jar:file:/tmp/model.zip”)
predictions = model.transform(data)
```

Scala - Write Bundle

```
def writeModel(bundlePath: String, model: PipelineModel, predictions: DataFrame) {
  val context = SparkBundleContext().withDataset(predictions)
  try {
    model.writeBundle.save(BundleFile(bundlePath))(context)
  } finally {
    bundle.close()
      }
  }
writeModel(“jar:file:/tmp/model.zip”, model)
```

Scala - Read as SparkBundle

```
import org.apache.spark.ml.Transformer
def readModel(bundlePath: String) : Transformer = {
  val bundle = BundleFile(bundlePath)
  try {
    bundle.loadMleapBundle.get.root
  } finally {
    bundle.close()
  }
}
val model = readModel("jar:file:/tmp/model.zip")
val predictions = model.transform(data)
```

Scala - Read as MLeapBundle

```
import ml.combust.mleap.runtime.frame.Transformer
def readModel(bundlePath: String) : Transformer = {
  val bundle = BundleFile(bundlePath)
  try {
    bundle.loadMleapBundle.get.root
  } finally {
    bundle.close()
  }
}
val model = readModel("jar:file:/tmp/model.zip")
val predictions = model.transform(data)
```

MLeap with MLflow examples

Python - Log model as bundle

```
 #  Log mleap model
mlflow.mleap.log_model(spark_model=model, sample_input=testData, \
    artifact_path="mleap-model")

# Log mleap schema file for MLeap runtime deserialization
schema_path = "schema.json"
with open(schema_path, 'w') as f:
  f.write(testData.schema.json())
mlflow.log_artifact(schema_path, "mleap-model")
```

Python - Read model as SparkBundle

```
from pyspark.ml import PipelineModel
from mleap.pyspark.spark_support import SimpleSparkSerializer

client = mlflow.tracking.MlflowClient()
run = client.get_run(run_id)
bundle_uri = f"file:{run.info.artifact_uri}/mleap-model/mleap/model"

model = PipelineModel.deserializeFromBundle(bundle_uri)
predictions = model.transform(data)
```

Scala - Read model as MLeapBundle

```
val schemaPath =
client.downloadArtifacts(opts.runId,"mleap-model/schema.json").getAbsolutePath
val schema = MLeapBundleUtils.readSchema(schemaPath)
val records = readData(opts.dataPath)
val data = DefaultLeapFrame(schema, records)
val modelPath =
client.downloadArtifacts(runId,"mleap-model/mleap/model").getAbsolutePath
val model = readModel(s"file:${modelPath}")
val transformed = model.transform(data).get
val predictions = transformed.select("prediction").get.dataset.map(p => p.getDouble(0))
```

MLeap Problems - Doc & Examples

- Very light and incomplete documentation
- Very few examples - only 4 notebooks updated 2 years ago
- None of which work out of the box
- All examples are Jupyter notebooks
- No normal Scala sbt and files example
- For Scala they use notebooks with Toree
  - https://github.com/combust/mleap-demo/blob/master/notebooks/airbnb-price-regression.ipynb
  - IMPORTANT!!! You may have to run this next block of code a few times to get it to work - this is due to another bug in Toree. For me, running it twice works.

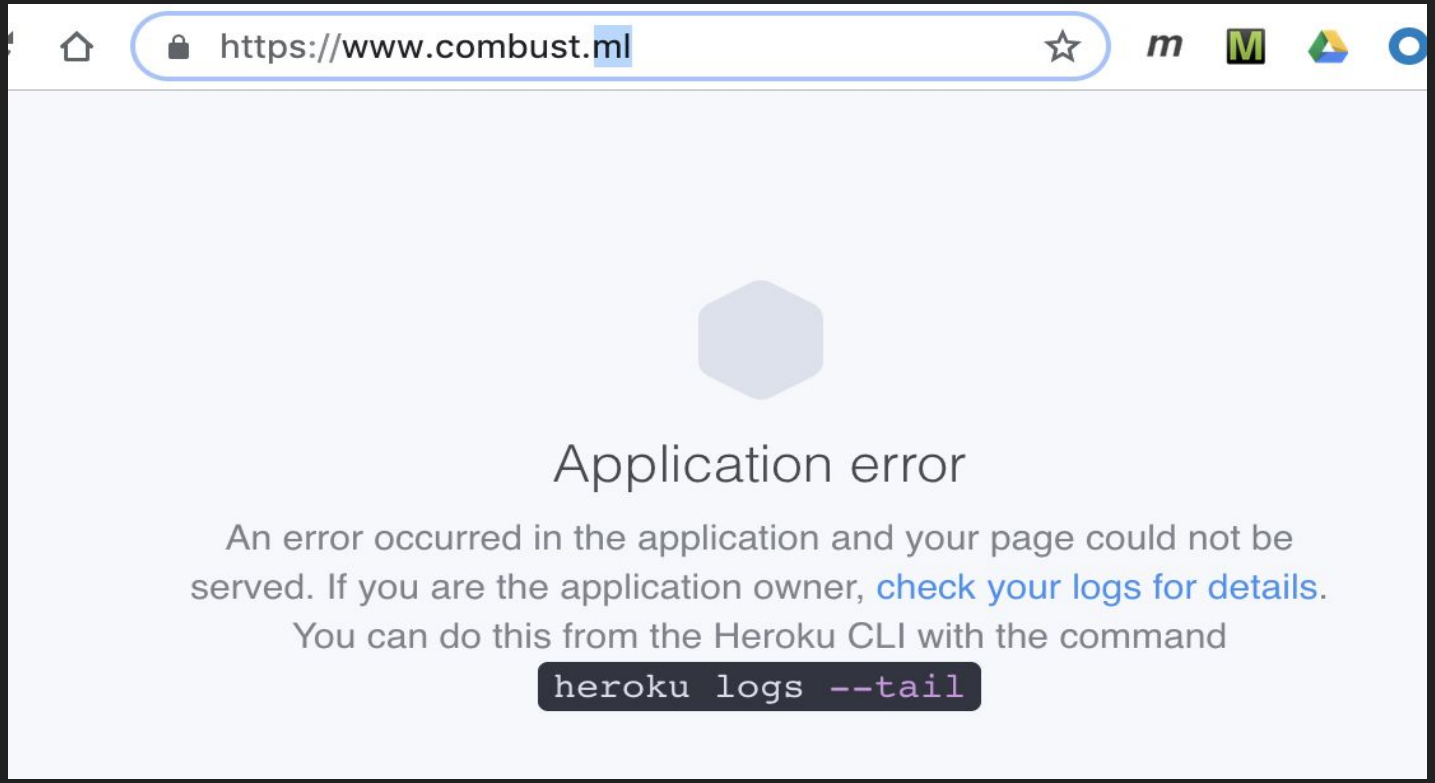
MLeap Problems - Error Handling

- Proper error handling is a must
- MLeap has poor error handling - too many NPEs
- If you specify bad bundle path you get an NPE and have no idea what went wrong
- Often silently will create a bad bundle (e.g. no bundle.json) and you only get the error when you try to deserialize

MLeap support and maintenance

- Combust was the company behind MLeap
- https://github.com/combust
- Combust has stopped support of MLeap
- Their web site is dead

Combust Home Page



MLflow SageMaker MLeap Container

```
mlflow sagemaker \
  build-and-push-container \
  --build \
  --no-push \
  --container sm-wine-pyspark-mleap

mlflow sagemaker run-local \
  --model-uri runs:/7e674524514846799310c41f10d6b99d/mleap-model \
  --port 5001 \
  --image sm-wine-pyspark-mleap
```

MLflow MLeap Java Scoring Server

```
git clone https://github.com/mlflow
cd mlflow/mlflow/java
mvn package
java -cp
scoring/target/mlflow-scoring-1.4.1-SNAPSHOT-with-dependencie
s.jar \
  org.mlflow.sagemaker.ScoringServer \
  models/mleap-model \
  5001
```

MLeap Sampler Code

- Sode examplesthat execute all different ways to read and write MLeap bundles.
- Dimensions:
  - With or without MLflow
  - Python or Scala
  - Score as SparkBundle or MLeapBundle (no Spark dependencies using just the MLeap Runtime).
- Code:
  - Python scripts: https://github.com/amesar/mleap-sampler
  - Databricks notebooks

Thank you

Have a good day