# Implement a Minimal Viable Data Product Using Deeplearning4J

This article demonstrates data flows and interaction between core components of a data-product. Our example use-case is automatic classification of images using neural networks.

After reading the article you will be able:

- to use DL4J in a Spark session / CDSW

- persist labeled training data in Kudu (in order to slice and dice the training set)

- execute a learned model in a Spark-shell session

- execute a learned model in a Spark-streaming job

This means: we go from learning to production, not perfectly robust, but end-2-end!

## Content

Scope 1

1. Ingest and convert raw images 3

2. Train a model from labeled images 3

3. Query for a specific training set 3

4. Variation of model parameters 3

5. Evaluation of model quality 3

6. Predict the class of unknown images 3

Conclusion & Outlook 4

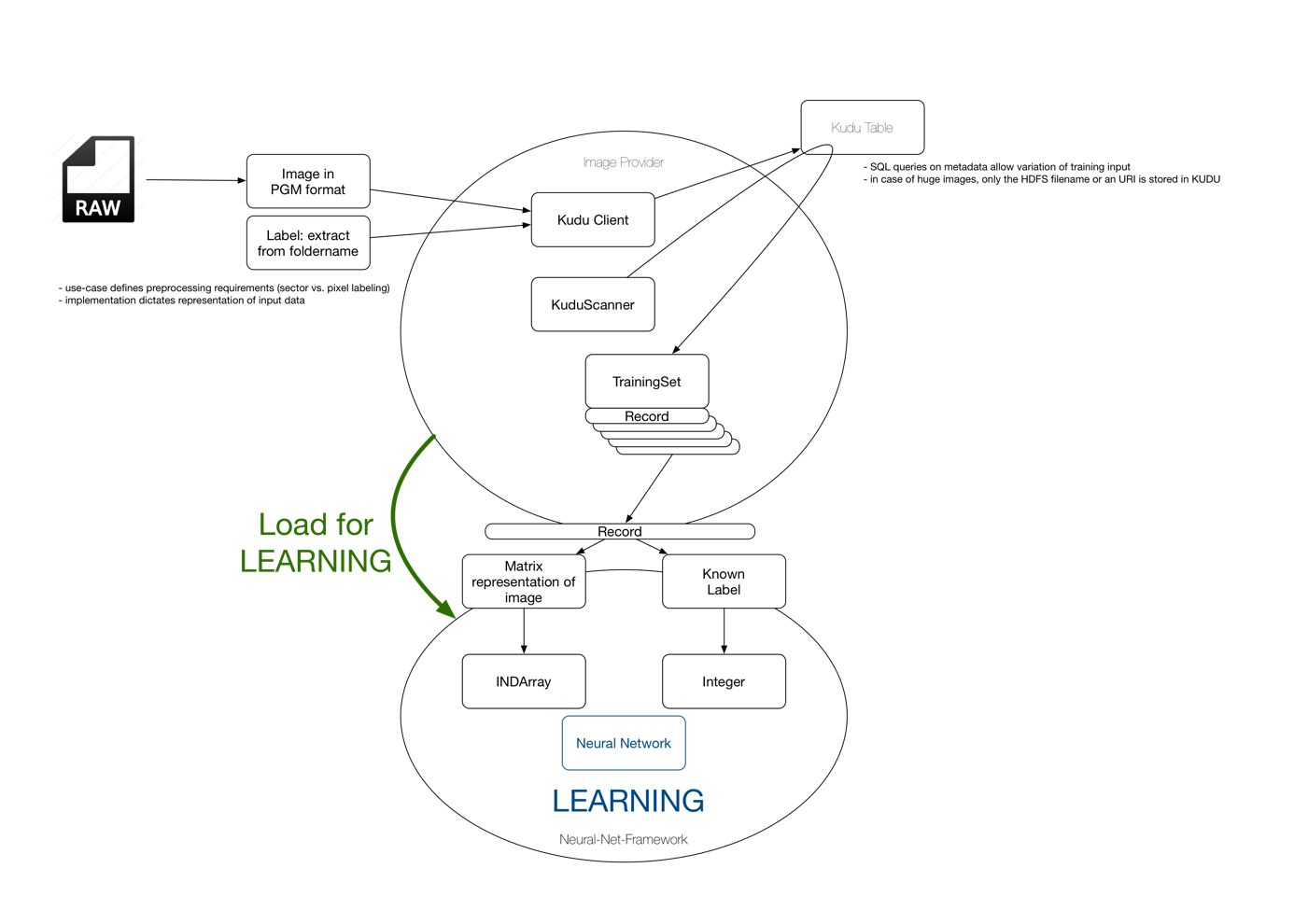
## 

## Scope

In order to solve the image classification problem, we have to act in multiple roles. Data engineering is needed in order to prepare data and a robust data pipeline. The input for our machine learning framework needs to be normalized and learning parameters have to chosen carefully – eventually automatically. Finally, after the data scientist job is done, the model has to be executed and maintained in a production environment - this is a great time for DevOps people.

We have seen that, in many cases only one single person plays all the three roles. But in a real world scenario in enterprise environments this isn’t a standard. People normally act in different contexts and multiple collaboration models for co-exist. In this article you will learn, how to implement a minimal viable data product using Apache Spark and Deeplearning4J on a CDH cluster.

Figure 1 illustrates our desired data flow which allows us to learn the unknown inherent structure of a labeled image dataset.



**Figure 1:** Overview of data flows and data type conversion during automatic image classification with Deeplearning4J.

Parameter-variation and model validation lead to multiple models represented by multiple neural networks. Furthermore, multiple network architectures can be trained and compared with each other.

In a robust data-product it is important to identify an appropriate model type and reasonable learning parameters which provide training in a reasonable amount of time. Model robustness has to be evaluated over time using new – so far un seen – input. And finally, continuous learning offers a reference model which can be compared with the current production model. This kind of stability analysis provides information about or even triggers model update requirements.

Persisting model metadata and metadata of raw images are key factors in the process of build a meaningful model in a cost efficient way. We use Apache Kudu as storage layer for this purpose. Running a learned model in production requires a direct embedding of the ML toolkit into your data pipelines. Using a Java based DL framework allows a direct integration inside Apache Flume and Spark Streaming jobs, as long as the trained model hasn’t to be parallelized over multiple nodes.

Model deployment has to be managed without any interruption to existing data flows. Since Apache Flume reloads the configuration file automatically regularly, we can simply offer the name or ID of the model which has to be loaded in the Flume configuration via Flume’s standard configuration mechanism. A more enterprise like approach could be based on a Flume source which receives a signal via Zookeeper as soon as the model needs to be changed. Reloading of the model from Kudu or HDFS is done fully transparently then.

Let’s go and implement the data product. This is our TODO-list:

1. Ingest and convert raw images
2. Train a model from labeled images
3. Query for a specific training set
4. Variation of model parameters
5. Evaluation of model quality
6. Predict the class of unknown images

## 1. Ingest and convert raw images

For our first step we use the simple Java program Step1.java and convert a batch of PNG files[[1]](#footnote-1) into PGM format[[2]](#footnote-2). The image encoding is encapsulated in a helper-class named ImageConverte.java. The current approach uses the convert command provided by ImageMagic. We generate a shell script and execute that using the ProcessBuilder class. Other options for image conversion are discussed in this article on StackOverflow[[3]](#footnote-3).

The Flume based solution uses an ImageInterceptor. You can find a Flume configuration stub file in the cfg folder of the project. Finally, the ingestion into Hadoop is accomplished by writing the transformed data entity - in our case a PGM version of the initial image into HDFS.  
  
Since such images are pretty small, one should consider an Avro file for image aggregation. HBase and Kudu tables can be considered as convenient approach as long as the image size less then 2MB (for HBase) or 64kB (for Kudu) respectively. The benefit of both technologies is the random access pattern both provide. Solr should only be considered to collect metadata for later data exploration. The raw image should always be references from a Solr document rather than being persisted in the index.

In order to initialize the Kudu image store you have to run the program Step2.java once.

## 2. Train a model from labeled images

Before we start learning multiple models, we prepare a model store using the Java program Step3.java.

There are two different implementations in our project. First, the ImageClassifierHDD, which simply persists the model files in a local folder on the workstation where the code is executes. We use this simply during development to test our network. You can run the Java program Step4 in order to create a simple model for character recognition.

The input data is provided by a DataSetIterator component. Deeplearning4J provides the MnistDataSetIterator out of the box. It allows accessing the binary MNIST database files.

DataSetIterator mnistTrain = **new** MnistDataSetIterator(*batchSize*, **true**, *rngSeed*);  
DataSetIterator mnistTest = **new** MnistDataSetIterator(*batchSize*, **false**, *rngSeed*);

If we chose to use a different data source – such as an HBase table or a Kudu table – we have to implement a custom DataSetIterator.

This brings us to the next step – but pretty essential part – of our data product. We need a reliable, but flexible way of training set preparation. Here we simply use SQL queries on image metadata. The idea can be generalized to HBase scans, Solr queries, or even SPARQL queries if an external triple store is used. The essence is: any kind of query or filter gives us a bunch of image IDs or file names which allows us to bring the data into Deeplearning4J via a DataSetIterator.

## 3. Query for a specific training set

In this step we provide a DataSetIterator for Kudu. The boundary conditions are simple: (1) we use a single machine for training the model. This means, the parallel query execution on Kudu leads to a single stream of input data. This can be used to define partitions for parallel learning at a later stage – one simple specifies a query per partition or batch.

The following link shows an example of a DataSetIterator which has to be used as a starting point for our own implementation:

<https://github.com/deeplearning4j/dl4j-examples/blob/master/dl4j-examples/src/main/java/org/deeplearning4j/examples/recurrent/word2vecsentiment/SentimentExampleIterator.java>

## 4. Variation of model parameters

Now it is time to repeat multiple training runs on the same data with variable learning parameters.

The program we prepared for parameter variation and producing multiple models is named

ParameterVariationExperiment1.

The logic inside the three exemplary loops is subject to parallelization. As long as a single model can easily be trained on one CPU we can use a Spark job to manage all possible parameter sets and start model training in parallel for each individual parameter set in one executor. The downside of this approach is that the training data needs to be loaded into each particular executor completely. But since we deal with a problem which needs multiple iterations on a huge data set, the only reasonable alternative would be a model chain which is trained in one executor. For small models this can be considered – but if the storage capacity required by one particular model is in the range of the executor’s heap size this is not possible any more.

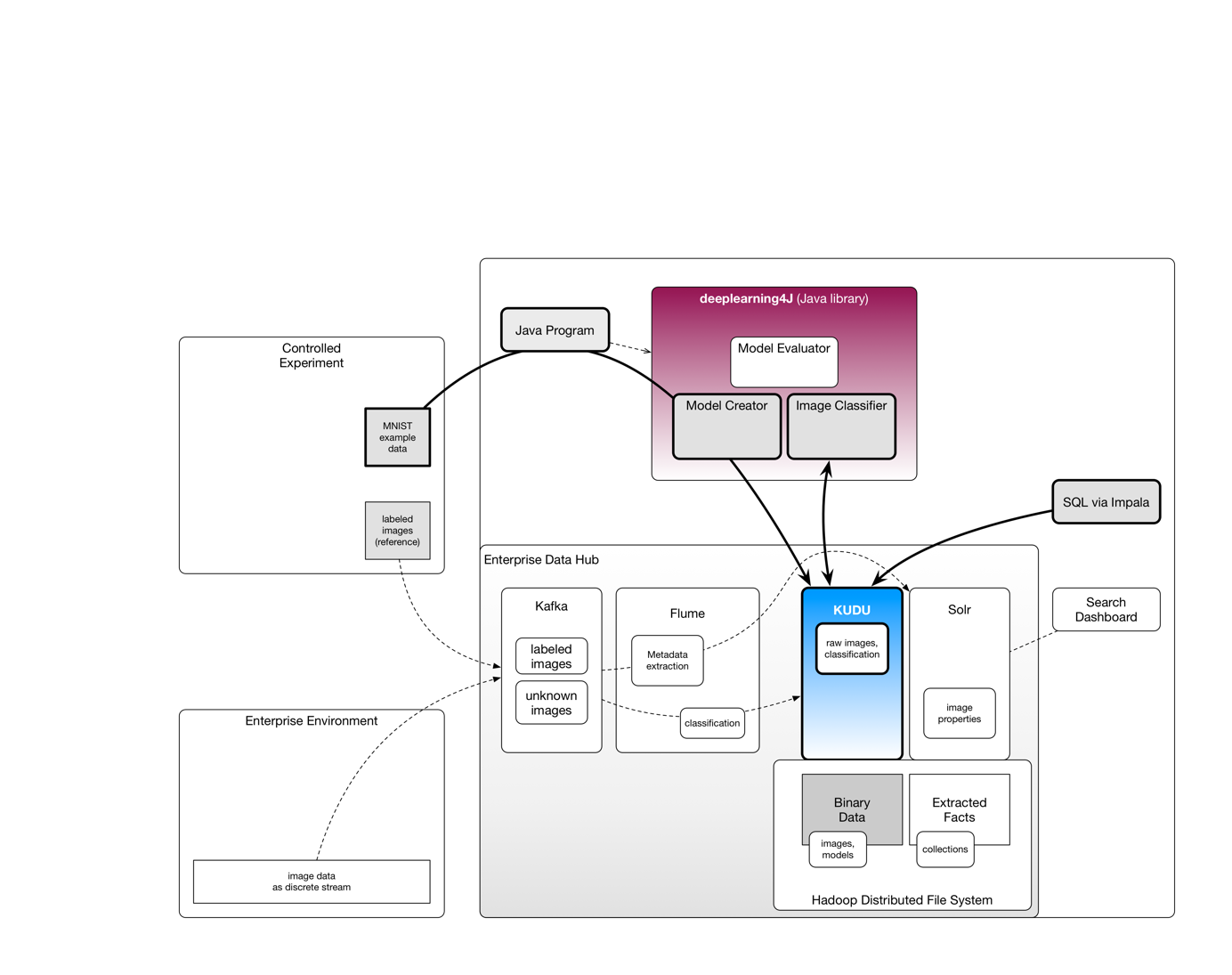
## 5. Evaluation of model quality

* Inspect model quality as a function of learning parametes and image preparation procedure
* This allows us to do an impact analysis on additional preprocessing

## 6. Predict the class of unknown images

* Java program with on particular image
* Flume Interceptor using the helper class and a SOLR sink
* Spark program with an RDD of images
* Spark streaming job Morphline or SolrJ client

Finally, we can draw a sketch to illustrate how the data flows through all the components and how they work together.



**Figure 2:** High level architecture and data flows for in a minimal viable data product using Deeplearning4J and CDH.

You will find the source code in this repository:

<https://github.com/kamir/cdsw-dl4j-mvdp-on-cdh>

In order to follow the examples, please prepare a 1 node quickstart VM which includes Apache Kudu as provided [here](https://kudu.apache.org/docs/quickstart.html).

Checkout our Github repository into the folder /home/cloudera inside the VM.

$ git clone https://github.com/kamir/cdsw-dl4j-mvdp-on-cdh

You can build the artifacts using:

$ mvn clean build

Now, if no error occurred, you can run the example by using the demo script:

$ bin/run\_demo.sh

## Conclusion & Outlook

This article illustrates a recurring pattern or simply: the anatomy of a data product. Multiple optimizations and variations exist. In many cases it is simply a different UI or a different source data set. Sometimes you need a specific model and the appropriate training algorithm which may not exist yet. But the components shown here can be seen as generalized architectural pattern – no matter which ML framework has been used.

Our next goal is to evaluate model parallelity and data parallel training, using Deeplearning4J on CDH.

1. PNG: <https://en.wikipedia.org/wiki/Portable_Network_Graphics> [↑](#footnote-ref-1)
2. PGM: <https://en.wikipedia.org/wiki/Netpbm_format> [↑](#footnote-ref-2)
3. ImageMagic: <http://stackoverflow.com/questions/5150503/image-magick-java> [↑](#footnote-ref-3)