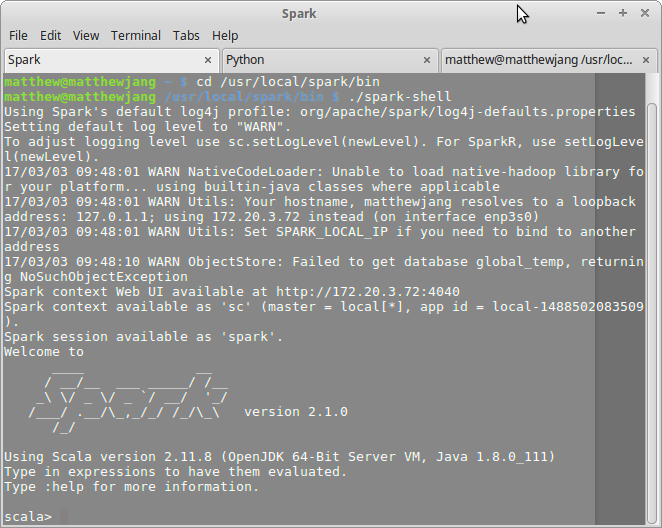
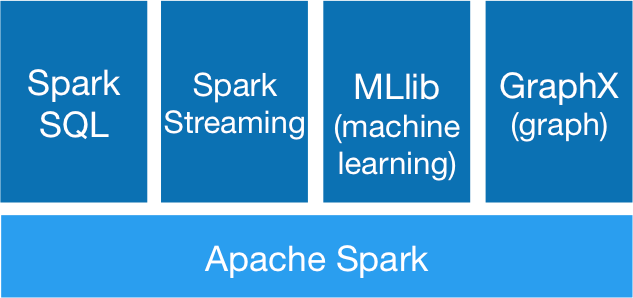
**Why Apache Spark?**

Apache Spark is an open-source cluster-computing framework developed by Matei Zaharia at the University of California, Berkeley’s AMPLab. Spark provides Scala, Java, and Python API (Application Programming Interface) created on a data structure called RDD (Resilient Distributed Dataset). Spark’s main feature is its in-memory cluster computing, which resulted in increased processing speed of an application, such as Hadoop’s MapReduce.

Here is what Spark looks like when run:

****

****Here is the general architecture of the Spark core:

As shown, Spark allows advanced analytics, such as Spark SQL, Spark Streaming, GraphX, and MLlib. I will be focusing on Spark streaming and MLlib.

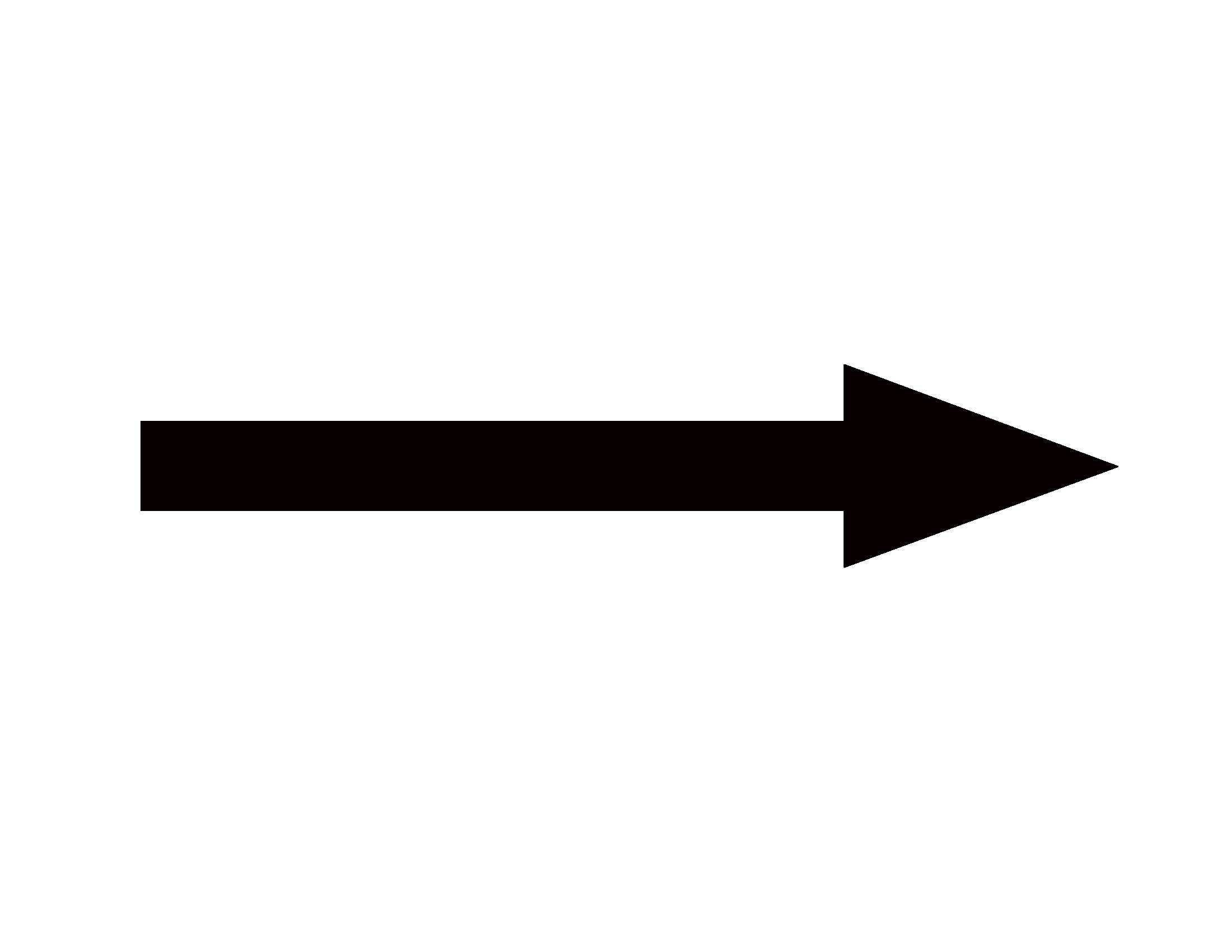
My idea and goal is to use the machine learning library of Spark to train the machine to recognize different types of malware (e.g. SQL Injection, Ransomware) that comes in through log files. MLlib is Spark’s machine learning library that consists of algorithms regarding classification, regression, clustering, collaborative filtering, dimensionality reduction, and other methods. MLlib consists of algorithms such as Naive Bayes, k-means clustering, and artificial neural networks (ANN).

So you might be wondering why we should even use machine learning at all. Well, it has its pros and cons, but I will focus on the benefits of machine learning.I believe we should use machine learning because of the following reasons:

1) Scale of data involved – human involvement will become infeasible as a company grows.

As Plura.io users increase, log data will also increase exponentially.

****

****

Log File

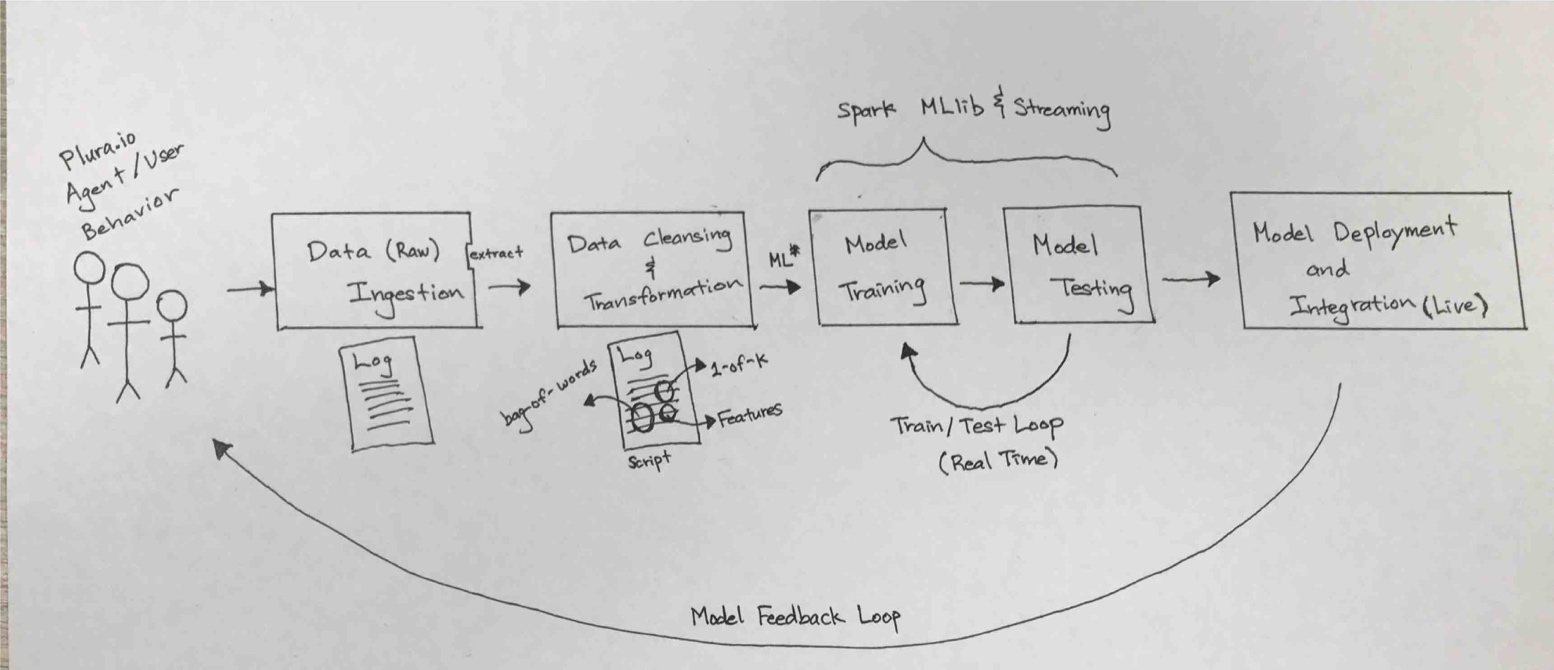
Big Data

2) Model-driven approaches (like machine learning) can benefit from uncovering patterns that humans cannot see.

3) Model-driven approaches (like machine learning) can avoid human emotional biases, which can lead to errors.

4) Can make company’s application/product personal to the user, and vice-versa.

Plura.io general machine learning pipeline/architecture:

****

Data will come in (“be ingested”) through the activities of the users. This data that has been ingested is called “raw data”, which is just data collected from a source. To use this raw data in MLlib, we need to extract it, “clean” it, and transform it into features/variables that are more amenable to machine learning. That is, raw data can be in the form of categories, when all machine learning models work on numerical representations as vectors. Thus, we must transform the raw data into numbers by built-in methods such as “1-of-k” or “bag-of-words”. Also, we can extract specific parts of a data that we want (e.g. IP Address, IP code) by writing a script.

The next step is where the meat of machine learning lies. We must now train and test our model continuously with our data at hand. That means we must use specific algorithms and models so that the machine finds patterns and makes predictions.