AMAZON MACHINE LEARNING AND ANALYTICS TOOLS

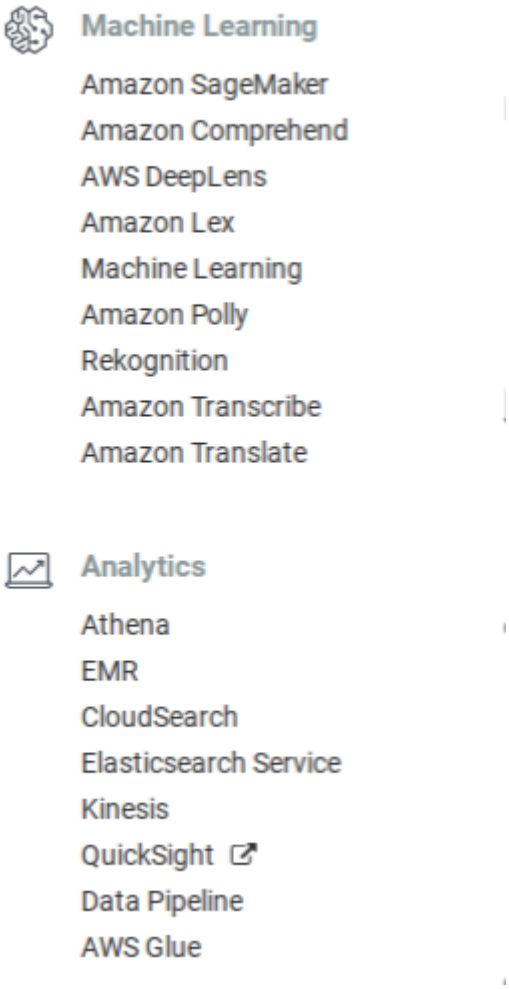
Here we begin our survey of Amazon AWS cloud analytics and big data tools. First we will give an overview of some of what is available. Then we will look at some of them in more detail in subsequent blog posts and provide examples of how to use them.

Amazon's approach to selling these cloud services is that these tools take some of the complexity out of developing ML predictive, classification models and neural networks. That is true, but could it be limiting.

In other words, linear and logistic regression and especially neural networks (used for deep learning) are not for the faint of heart. Amazon ML picks the algorithm for the user. Data scientists are used to being able to do that and to modifying the parameters for said model. Amazon would say users do not need to do that, as their algorithms will do what the data scientist does, which is to change the parameters automatically to reduce the error rate (i.e., the difference between predicted and observed values) to their lowest value.

The Amazon Machine Learning and Analytics Console

When you log into the Amazon AWS console it presents this list of items. You can pick from these and add them to your account. But be careful as the meter starts running when you do that.



The Machine Learning tools at the top are mainly geared toward voice, image, and text analysis, except for the Machine Learning model. (We already wrote about how to use Google Natural Language API here.)

Voice and image recognition is not really a business application. It is a service that you can program yourself using neural networks and train that with publicly downloadable free datasets or pay Amazon to use their cloud. But most people doing analytics in their daily jobs are not going to be interested in creating their own Apple Siri or other voice recognition type system.

Instead they are more like to benefit from less esoteric tools like Athena or QuickSight. QuickSight, in fact, is so easy the end user could use it. So you could set that up and let your end users play with it, thus freeing up your data science resources, somewhat.

Athena lets you wrap a schema around any data in Amazon S3 (i.e., one of their cloud storage products) and then run queries against that.

So let's look at a couple of these products briefly to see which might be of use to you.

First, a word. Anyone who uses Amazon EC2 knows that subscription fees can mount quickly. Amazon says that for most of the products listed on the AWS management console there are no up-front fees and you pay as you go. But note that the tutorials are not including in the free tier pricing, so watch your subscription fees. (You can create billing alerts on your Amazon account. So here would be a good place to stop and do that.)

Amazon Product	Overview

Machine Learning	ML does what a Python or Scala programmer using Spark or similar language and platform would do using Spark ML, TensorFlow, or other API. To use it, the user uploads data to Amazon S3 or Redshift. Then Amazon splits that into testing (30%) and training (70%) data sets. Then it builds predictive models and shows the results without requiring coding. But you do need to write a program to put your data into a format that ML can understand. In other to use the Amazon ML models you create what data science programmers would call a **feature-label** matrix. You do this by putting some feature that you want to predict (i.e., the dependent variable) and labels (i.e., independent variables) into a file. In other words, you might assume that sales are a function of price, time of year, advertising budget etc. Then Amazon ML will pick the best model to run predictions or your sales given the price, time of the year, advertising budget, etc. **Best** means the model that produces the most accurate results. To put that into data science terms, it will run logistic or linear regression against the data. But it does this automatically without you having to write any code.
Athena	This lets you wrap a schema around data loaded into Amazon S3 and run queries against it.
Elasticsearch Service	This is really just a cloud way of running ElasticSearch. So it is infrastructure and not a product. You would probably spend less money by installing your own system as all you need are virtual machines. Beside you will have to configure logstash and filebeats and the other connectors yourself and connect those to your applications and infrastructure systems, like web servers, application servers, firewall logs, and security detection tools. ElasticSearch is usually called **ELK**, for **ElasticSearch**, **Kibana**, and **LogStash**, as these 3 products are designed to work together. They are the most popular tool for gathering log data across an enterprise.
SageMaker	SageMaker uses notebooks. iPython (now called Jupyter) and Zeppelin are notebooks that data scientist have long used. These are interactive web pages where you can write Scala or Python or other code to query Spark and other data stores. Then you can hide that code and publish it is live web pages that your users can view.

DeepLens

Is an actual physical video camera and cloud service to do image recognition. Connects to SageMaker and other Amazon tools like Amazon Kinesis Video Streams.