H2O- The Next Big Things

Abstract:

How this pattern recognition technology can enable insights to be derived from data. It's really compelling because before machine learning you needed people to come up with creative ideas and essentially superimpose their own perspective on to a data set. Whereas now with machine learning, you can, essentially, get some insights into the different patterns that exist in the data.

You're basically facilitating that first step of a human being working with the information to try to understand what the segments are, maybe what they mean and give some direction to where that person goes and try to better analyze the data and then come up with some ideas

There are several distributed machine learning platforms which are trying to mitigate this effort. In this talk we will focus on H2O. H2O is an open-source distributed math-engine providing tuned Machine Learning library. H2O is for data scientists and business analysts who need scalable and fast machine learning. Unlike traditional analytics tools, H2O provides a combination of extraordinary math and high performance parallel processing with unrivaled ease of use.

Introduction:

Data is nothing but capturing events, day-to-day events of life and the nature of the universe, so that the fundamental characteristics of data reflect the universe, which is very scattered, and life, which is messy. Data is scattered and messy. These days, businesses are producing more data than ever before. But data, by itself is not interesting. You want to understand what are the underlying models and patterns behind it, so you can answer questions:

What would my customer do next? Where is he coming from? What was his journey like? Can I categorize this person's experience as good, bad, ugly? How do I make him happier?

Machine learning algorithms are the keys to creating models of what happened in the past, so you can use new data to predict what happens next. Machine learning is nothing new; the field has been around for decades. The heart of machine learning going into businesses is everyone wants to transform their particular industry, whether it's a healthcare hospital trying to improve patient treatment, whether it's an insurance company trying to predict and replace an underwriter, or whether it's a very large cargo company trying to figure out how to manage their fleets better. All of these guys are applying strategy to the problem and the notion that you can apply algorithms to make predictions better.

Amazon Web Services turned a lot of heads recently when it launched a machine learning platform aimed at making predictive analytics applications easy to build and run, joining cloud juggernauts Microsoft and Google with similar ML offerings. It turns out the cloud is very well-suited for this critical type of big data workload.

Previously the only way to do advances analytics was to buy an expensive stats package
like SAS or IBM's SPSS or use the emerging open source tools like R.

H2o is an open source machine learning platform and provide a combination of extraordinary math and high performance parallel processing with ease of use. H2O really makes it easy to get great results with minimal effort due to its user-friendly web interface, its high performance and scalability, and the built-in automation and model tuning options. H20 user can easily explore and model big data from MS Excel and RStidio and connect it with HDFS, SL, SQL, NOSQL data source. Algorithm include distributed tree and Regression such as gradient busting machine (GBM), random forest (RF), generalize linear modeling (GLM), k-means and principal component analysis (PCA).

It can work with HDFS on Hadoop environment. Thus, when working on HDFS, H2O makes Hadoop do math and analysis. It can be considered as a math engine that brings interactivity and scalability to big

data modeling. H2O allows large data set to be used in real time without the need of sampling. It is a better prediction analytical engine in Big Data Science.

Oxdata (http://h2o.ai/) provides software to allow data scientists to quickly and easily run machine learning models at scale. The intended audience seems to be people familiar with R who are limited by the scalability of R. Using H2O allows data scientists to distribute machine learning algorithms over a cluster. Not all machine learning algorithms are currently supported but the supported list is quite impressive. Many R functions and structures are supported but this will never likely be a clone of R. Oxdata seems to use a freemium model: the basic software is free and open source. Enterprises can choose to buy a premium license that provides them with 24/7 support, help with optimizing and scaling clusters, etc.

INTRO OT H2O:

Erin LeDell, Data Scientist, H2O told “Data Science consists of three major steps: **Problem Formulation**, **Data Processing** and **Machine Learning**.”

With help of H2O we can make better predictions. Harness sophisticated, ready-to-use algorithms and the processing power you need to analyze bigger data sets, more models, and more variables.

Get started with minimal effort and investment. H2O is an extensible open source platform that offers the most pragmatic way to put big data to work for your business. With H2O, you can work with your existing languages and tools.

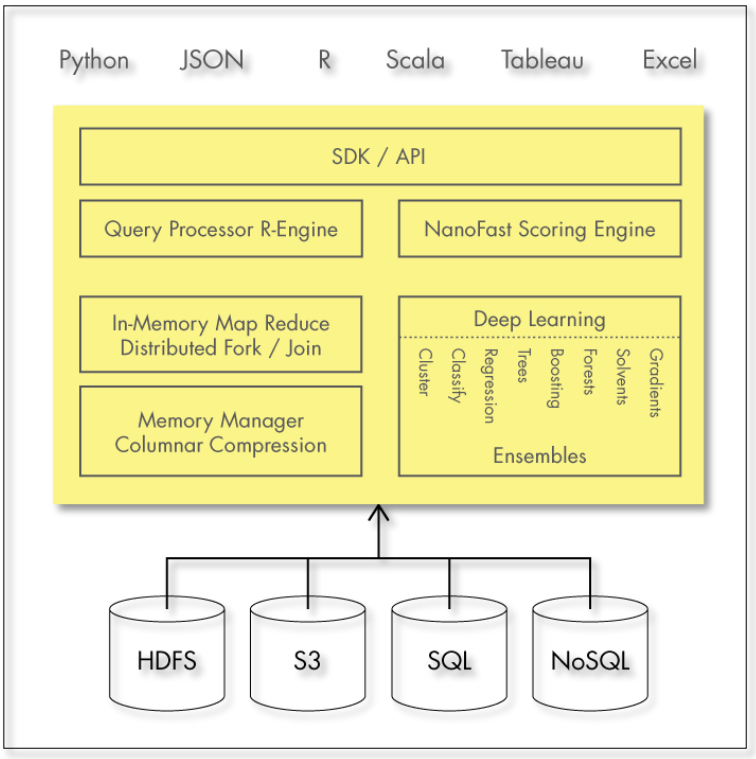
H2O does in-memory analytics on clusters with distributed parallelized state-of-the-art Machine Learning algorithms. It operates at layer level and is coded to at different layers to approach different tasks and problems.

The rationale for H2O can be summarized by:

- Faster: Minutes vs. hours/days
- Bigger: Bigger dataset / Cluster Mode
- Better: Ease of Sampling and Feature Selection

H2O also features native support for Java, Scala, and Python. The solution’s interface is driven by JSON APIs, which makes it easy to plug into your organization’s existing tools and processes to train your data and continuously improve your models and predictive accuracy.

H2O- The Next Big Things



Combine the power of highly advanced algorithms, the freedom of open source, and the capacity of truly scalable in-memory processing for big data on one or many nodes. These capabilities make it faster, easier, and more cost effective to harness big data to maximum benefit for the business. H2O makes it fast and easy to derive insights from your data through faster and better predictive modeling. Existing Big Data stacks are batch oriented. Search and analytics need to be interactive. Use machines to learn machine-generated data. And more data beats better algorithms.

- H2O does in-memory analytics on clusters with distributed parallelized state-of-the-art Machine Learning algorithms. It operates at layer level and is coded to at different layers to approach different tasks and problems.
- In-memory distributed K/V store layer: It is similar to the Java memory Model but this is also distributed. Both read and write are fully cacheable. All reads and writes go through a non-blocking hash table. H2O keeps all data in heap. The advantages of keeping all data in heap is that it is as fast as possible and is easy to code (pure Java). When Java heap gets too full, H2O does user-mode swap-to-disk.
- Pre-baked Algorithms layer: Variety of fully optimized and fully featured algorithms in H2O. Within each algorithm, H2O supports full range of tuning parameters and options.

H2O- The Next Big Things

H2O software stack:

The top section shows some of the different REST API clients that exist for H2O.

The bottom section shows different components that run within an H2O JVM process.

All REST API clients communicate with H2O over a socket connection.

The embedded H2O Web UI is written in JavaScript.

R scripts can use the H2O R package ['library(h2o)']. Users can write their own R functions than run on H2O with 'apply' or 'ddply'.

Python scripts currently must use the REST API directly. An H2O client API for python is planned

An H2O worksheet for Microsoft Excel is available. It allows you to import big datasets into H2O and run algorithms like GLM directly from Excel.

Users can pull results from H2O for visualization in Tableau.

JVM Components:

An H2O cloud consists of one or more nodes. Each node is a single JVM process. Each JVM process is split into three layers: language, algorithms, and core infrastructure.

H2O- The Next Big Things

The language layer consists of an expression evaluation engine for R and the Shalala Scala layer. The R evaluation layer is a slave to the R REST client front-end. The Scala layer, however, is a first-class citizen in which you can write native programs and algorithms that use H2O.

Underneath the covers, the H2O JVM sits on an in-memory, non-persistent key-value (KV) store that uses a distributed JAVA memory model. The KV store holds state information, all results and the big data itself. H2O keeps the data in a heap. When the heap gets full, i.e. when you are working with more data than physical DRAM, H20 swaps to disk. The main point here is that the data is not in R. R only has a pointer to the data, an S4 object containing the IP address, port and key name for the data sitting in H2O. The R H2O package communicates with the H2O JVM over a REST API. R sends RCurl commands and H2O sends back JSON responses. Data ingestion, however, does not happen via the REST API. Rather, an R user calls a function that causes the data to be directly parsed into the H2O KV store.

Conclusion:

H2O provides data mining algorithms that are highly efficient. You can interface with H2O with the several APIs such as their R API. The benefit of combining R and H2O is that it H2O is very good at exploiting multi cores or clusters with minimal effort of the user. It is much harder to achieve the same efficiency in R.

The reason why H2O is much faster is that they have a very good indexing of their data and their algorithms are written such that they exploit parallelism to the fullest

R with the default matrix dynamic libraries can only use one CPU core. Revolution R community edition ships with the Intel Math Kernel Library. This allows to do some matrix computations in parallel but definitely not as efficient as H2O. For SAS it is a bit harder to say anything considering its closed source but based on my CPU utilization I would assume that they have a similar approach as Revolution R. Their matrix algebra exploits parallelism but they algorithms are not as efficient as H2O. Their data storage is also not as efficient as H2O.

Lastly, H2O with R comes at a very different price tag than SAS.

RESOURCES:

- H2O.ai - Fast Scalable Machine Learning:
  http://h2o.ai/product/

- H20 world: Learning cab be fun:
  https://www.youtube.com/user/0xdata

- H2O for Deep Learning, evolution of High Performance Computing (HPC) and the future of HPC:
  http://www.kdnuggets.com/2015/01/interview-arno-candel-h20-deep-learning.html