SKYTREE®
BIG DATA ANALYTICS

WHITE PAPER

Analyzing Massive Datasets

BY ALEXANDER GRAY PhD, CTO, SKYTREE

Machine Learning (ML) is the modern science of discovering patterns and making predictions from complex data.

FOUR TAKE-AWAYS

In this brief white paper, we'll answer four questions that are arising ever more frequently, as interest in "analytics" on massive datasets appears to be surging in virtually all industries and fields. In particular, we'll attempt to bridge the large gap between the world of academic research in statistical methods and algorithms and the world of commercial offerings in the analytics space.

1. The difference between simple and advanced analytics
2. How domain-specific is advanced analytics?
3. Can existing commercial solutions really handle massive data?
4. Can state-of-the-art advanced analytics be done on massive data?

ABOUT THE AUTHOR

Alexander Gray PhD, CTO of Skytree, received a bachelor's degrees in Applied Mathematics and Computer Science from the University of California, Berkeley and a PhD in Computer Science from Carnegie Mellon University. He began working with massive scientific datasets in 1993 (long before the current fashionable talk of "big data") at NASA's Jet Propulsion Laboratory in its Machine Learning Systems Group. High-profile applications of his large-scale ML algorithms have been described in staff-written articles in Science and Nature, including contributions to work selected by Science as the Top Scientific Breakthrough of 2003. He has won or been nominated for a number of best paper awards in statistics and data mining and is a recipient of the National Science Foundation CAREER Award in 2009. He gives invited tutorial lectures on massive-scale data analysis at the top data analysis research conferences, government agencies, and corporations, and is a member of the prestigious National Academy of Sciences Committee on the Analysis of Massive Data. He is currently a professor in the College of Computing at Georgia Tech.

1 WHAT IS ANALYTICS?

There are many things lumped under the heading of "analytics", some simple, and some sophisticated. In general, we make the following distinction:

Simple analytics. Some basic but quite useful statistics can go a long way. These include things like counts, averages, and histograms. Most of this can be formulated in terms of simple SQL queries to a large relational database or data warehouse. An example might be "What is the average amount of grocery sales in each region?" The computational problem boils down to computing a set of SQL queries quickly. Straightforward data parallelism, such as that provided by Map Reduce/Hadoop, can in principle speed up such operations by a factor of p, given p processors. (In practice there is significant overhead preventing the full theoretical speedup.) Vendors of solutions include Aster, Greenplum, and Netezza.

Advanced analytics. This includes things like discovering patterns in data and predicting quantities based on values of other quantities. This covers the academic fields of machine learning (ML), data mining, pattern recognition, multivariate statistics, and others (we will just use the terms "machine learning" or "statistics" to refer to them all, as there is no major conceptual difference between these research communities). The leap in sophistication from simple analytics to such techniques is generally large, in terms of both the statistical background needed by users to get the most out of employing such techniques, and (especially) the computational background needed to make such techniques possible on larger than modest datasets.

However, the payoff is often large, or game-changing. Parts of virtually every scientific or engineering field have already been revolutionized by such techniques in the last 20 years or so, such as what is now called computational biology. Penetration into industry has only just begun, with some notable high-profile exceptions such as the US post office's ML-based automatic zipcode recognition, Google's ML-based ad serving, Netflix's ML-based movie recommendations, the ML-based speech recognition in most help lines, the ML-based fraud detection schemes used by credit cards, and numerous others.

> **Main take-away point:** *Advanced analytics and SQL querying are entirely different beasts.*

2 HOW DOMAIN-SPECIFIC IS ADVANCED ANALYTICS?

In other words, is machine learning for business problems very different from machine learning for astronomy problems? We have worked on a diverse range of data analysis projects spanning almost two decades, the answer is "yes and no".

Domain knowledge can be incredibly important in the identification of the statistical issues and the most appropriate methods for the problem/goals at hand. There are entire fields such as geostatistics, astrostatistics, or econometrics, which focus on these issues. However, in the end, the final best ML method needed is often a general one that can be applied in many fields (such as kriging or Gaussian process regression). Sometimes there are complications that deviate from the "textbook" situation. However, even most of these have been identified and studied as general statistical issues that can occur in many fields. Examples include systematic or non-systematic outliers or noise, missing values, different costs for false positives versus false negatives, truncated values (called "censoring"), or the training and test sets being drawn from different distributions (called "covariate shift" or "concept drift"). Sometimes certain problems legitimately contain aspects or constraints that are unique, or not exactly like what is found in any other domains. An example might be certain structure in the variables (such as spatio-temporal relationships, or known relationships

between the variables) that is informed by prior knowledge. Such problems can benefit from a truly custom-designed model, which requires full statistical expertise. Note that certain other situations suggest custom modeling, such as very small data sizes and certain interpretability constraints.

Nonetheless, in many if not most cases, a toolkit of state-of-the-art generic building blocks corresponding to well-defined standard statistical tasks (such as classification, regression, etc.) solves the problem well, as long as the toolkit is comprehensive enough. Thus a set of the best existing general machine learning methods, possibly in conjunction with the statistical expertise needed to handle any "complications" as mentioned, can solve many if not most problems well across many fields. Note that the state of the art machine learning methodology evolves rapidly and is typically years if not decades ahead of what is known or used in various domains. A true machine learning expert with guidance in the goals and issues of the domain area will always do better than a domain expert with light machine learning knowledge. There are domain-specific issues, but these tend to be more "nuts-and-bolts" issues rather than abstract ones that change the mathematical approach needed.

> **Main take-away point:** *State-of-the-art ML methods and statistical methodology is very "horizontal" (generic across fields). The "vertical" (domain-specific) part consists of data capture and preprocessing, software integration, and less frequently, custom modeling.*

3 CAN EXISTING COMMERCIAL SOLUTIONS REALLY HANDLE MASSIVE DATA?

All of the offerings today creating buzz are actually meant for simple analytics (SQL queries). For certain types of problems, some of these solutions can indeed provide large speedups over previous solutions. However, nothing really exists to make most of advanced analytics scalable. Advanced analytics methods can be found in well-known packages including Excel, Matlab, R, SAS, SPSS, and others. These generally feature some suite of methods, of varying sizes. (As a side note, none comes close to being able to claim it is a comprehensive collection of the true state-of-the-art in ML methods.) Advanced analytics is slow in all of these solutions. There is no easy fix for these solutions, as they were not fundamentally designed for scalability from the ground up. Solutions based on different computer science are needed to make advance analytics tractable on massive datasets.

What about databases and data warehouses? Though some products offer in-database advanced analytics, they utilize straightforward implementations, aside from addressing one aspect of scalability well, which is out-of-core processing. Therefore, only the very simplest few of the advanced analytics methods offered by these products can be claimed to be at all scalable on massive datasets.

Maybe vertical store is the answer? Vertical store is appropriate for simple analytics, where the important observation that many common SQL queries involve only one or a few fields was made. This notion is too simple to treat the more complex fundamental operations needed in advanced analytics.

Doesn't Hadoop solve everything? Map Reduce/Hadoop was originally designed to easily parallelize text processing tasks, such as creating search indices. The class of problems for which it is applicable consists of such "embarrassingly parallel" tasks. As in the case of RDBMSs above, only the very simplest few of the advanced analytics methods is a good fit for this class. Beyond these, advanced analytics methods are generally not embarrassingly parallel tasks.

What about hardware approaches to parallelism? GPUs? Appliances? Specialized devices such as graphics processing units (GPUs) and appliances such as Netezza's have, again, been applied to some of the simplest advanced analytics methods that admit embarrassingly parallel solutions.

> **Main take-away point:** *Commerical solutions do net yet exist which provide a suite of state of the art ML methods that can scale to massive datasets.*

4 CAN STATE-OF-THE-ART ADVANCED ANALYTICS BE DONE ON MASSIVE DATA?

Due to the importance of this topic across virtually all fields in the foreseeable future, a committee commissioned by the prestigious National Academy of Sciences (of which the author is a member) is developing a report detailing the current state of affairs and its recommendations for further research. The research community has increasingly studied various aspects of the problem of how to scale ML methods to massive data. Work in this general area has been slow and is fairly dispersed, mirroring the dispersed nature of the statistical sciences as well as the diffculty of simultaneously achieving sophistication in both statistical and computational disciplines. For certain ML methods, no great solutions exist yet. But for many, certain research groups have developed powerful new algorithms in recent years. For example, Georgia Tech's FASTlab (www.fast-lab.org), the only academic lab devoted to the problem comprehensively, has developed a number of the current most effective techniques, and keeps close tabs on developments in the field by other researchers.

A NEW APPROACH

At Skytree, we believe a radically different approach is required in order to make advanced analytics tractable on massive datasets, based on three things:

New fast algorithms. Efficient algorithms can make all of the advanced analytics methods 10–10,000x faster, on a single machine.

New data representations. Just as relational databases are based on B-trees for speed, and search engines are based on inverted indices for speed, advanced analytics can utilize specialized data representations to accelerate its operations.

New distributed systems. Parallel/distributed approaches which are specialized for advanced analytics tasks can tackle the many ML methods which are not simply embarrassingly parallel.

SKYTREE®
BIG DATA ANALYTICS

www.skytreecorp.com
For more information: info@skytreecorp.com