Skytree: A Scalable Data Science Environment for Massive Datasets

Nicholas M. Ball, Alexander Gray
Skytree, Inc., 1731 Technology Drive, Suite 700, San Jose, CA 95110
http://www.skytree.net nick@skytree.net

Introduction

Skytree is the world's most advanced machine learning software. It acts as a machine learning server to allow advanced data mining on large data (**Figure 1**), e.g., within one's data processing pipeline, or more specialized science project. Skytree's Alex Gray also headed the FASTlab group at the Georgia Institute of Technology. The group holds several records for the fastest implementations of well-known machine learning algorithms. Algorithms that otherwise scale as, e.g., N^2, for N objects, are implemented to scale linearly, without loss of accuracy. While each specific use case will remain problem-driven, the underlying tools are not dataset-specific. Thus, the installation of Skytree on the one's computing infrastructure makes possible the practical use of these algorithms by users who are not data mining specialists. The software quality and robustness renders it suitable for both its primary function: enterprise business use, and publication-quality research.

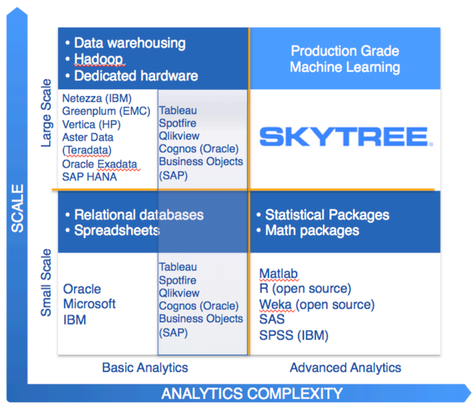

Figure 1: Skytree allows both advanced analytics, and its application to large data. This fills a vacant parameter space that is essential for future big data analysis.

SKYTREE®
BIG DATA ANALYTICS

Why Skytree?

Skytree instantiates 3 fundamental differentiators (**Figure 2**) that make its value unique: **(1) Breadth of methods:** there is no single best machine learning algorithm, so best results require multiple methods; **(2) Speed and scalability:** without Skytree's NlogN or better scaling, advanced machine learning on large datasets is intractable; **(3) Ease of use:** Skytree enables cutting-edge, complete, advanced analyses in a single command.

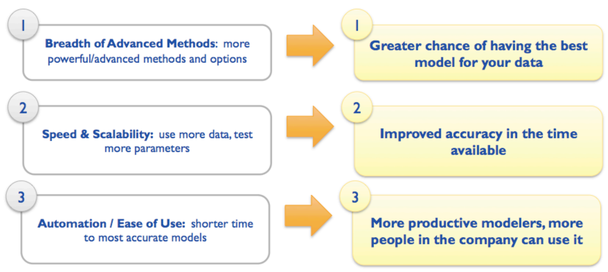

Figure 2: Skytree's key differentiators compared to other machine learning software

Using Skytree

Skytree supports multiple skill levels of users (**Figure 3**): business analysts via GUI, intermediate users via an API, and advanced users via the command line. The full power is available to all users.

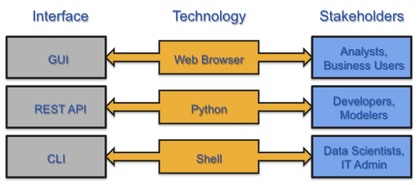

Figure 3: Multiple user skill levels are supported

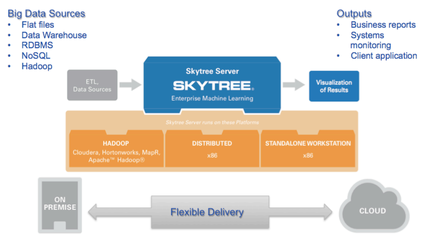

Figure 4: Skytree deployment options

Skytree is an x86 Linux executable that can run on any environment that runs Linux software. Hence, standalone, distributed clusters, and multiple Hadoop distributions are all supported.

Data Science Examples

Skytree's data science projects are problem-driven solutions of value to customers, utilizing the power of Skytree's machine learning algorithms. We also carry out proof-of-concept and long-term engineering projects to advance the state-of-the art in scalable machine learning on massive datasets. The examples shown here represent a cross-section of our data science work.

NlogN Scaling

We demonstrate our claim to NlogN (or better) scaling on real data on a real system using the 470,991,651 row dataset 2MASS (2 Micron All-Sky Survey, [4]) of astronomical objects. **Figure 5** shows linear scaling as the data size increases for the nearest neighbors algorithm [1]. Normally, this algorithm scales as N^2.

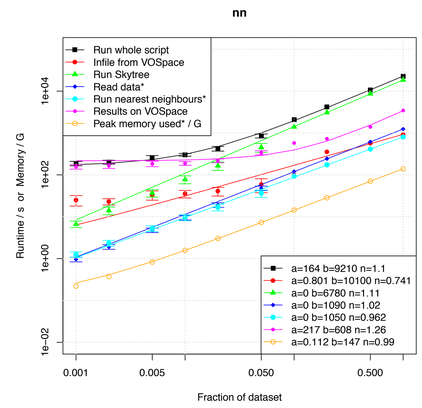

Figure 5: Skytree scaling for 470,991,651 objects in the 2MASS catalog. A naïve implementation would scale as N^2; Skytree's scales as N

Outliers

We utilize Skytree's K-means, kernel density estimation, and nearest neighbor algorithms to locate outliers [2,3] within the same 2MASS dataset. **Figure 6** shows a visualization.

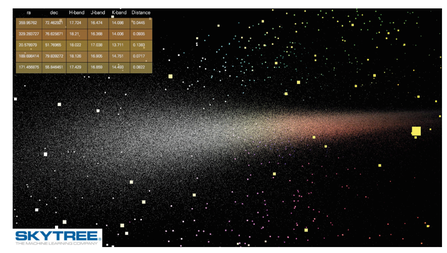

Figure 6: Visualization of outliers in the 2MASS dataset

Customers

As an enterprise software company, Skytree has multiple commercial customers with whom we have demonstrated quantitative success and dollar value to them. These include (**Figure 7**) eHarmony for algorithmic pricing, Brookfield for property valuation, Adconion for web advertising, the US Golf Association, and the SETI Institute (Search for Extraterrestrial Intelligence).

Figure 7: Some of Skytree's customers

Benchmarks & Strong Scaling

Figure 8 shows typical performance benchmarks versus other data mining implementations, and an example of *strong scaling*, the proportional increase in computing speed with more computing nodes on a cluster.

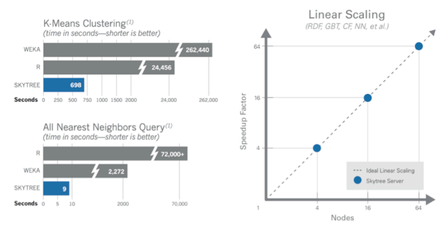

Figure 8: Skytree benchmarks and strong scaling

Weak Scaling

Figure 9 shows weak scaling, the ability to distribute a large dataset across a computing cluster, utilizing more memory than any individual machine has available.

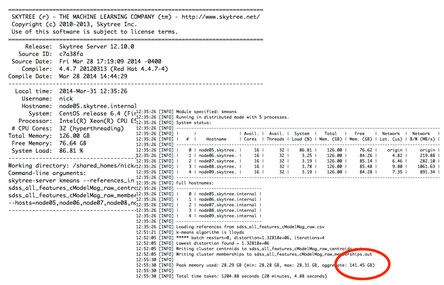

Figure 9: Skytree weak scaling

References

1. Ball, N.M., et al., 2014, *CANFAR+Skytree: The World's First Cloud Computing Data Mining System for Astronomy*, in preparation
2. Chandola, V., Banerjee, A., Kumar, V., 2009, *Anomaly Detection: A Survey*, ACM Computing Surveys (CSUR), 41(3) 15
3. Hodge, V.J., Austin, J, 2004, *A Survey of Outlier Detection Methodologies*, Artificial Intelligence Review, 22(2) 85
4. Skrutskie, R.M., et al., 2006, *The Two Micron All Sky Survey (2MASS)*, AJ 131 1163