REVOLUTION ANALYTICS EXECUTIVE BRIEFING

The Rise of Big Data Spurs a Revolution in Big Analytics

By Norman H. Nie, CEO Revolution Analytics

The enormous growth in the amount of data that the global economy now generates has been well documented, but the magnitude of its potential impact to drive competitive advantage has not. It is my hope that this briefing urges all stakeholders—executives who must fund analytics initiatives, IT teams that support them and data scientists, who uncover and communicate meaningful insight—to go boldly in the direction of "Big Analytics." This opportunity is enormous and without precedent.

Growth in Data Volumes

"Big Data" refers to datasets that are terabytes to petabytes (and even exabytes) in size and comes from various sources, including mobile digital data creation devices and digital devices used in both the public and private sectors. The massive size of Big Data go beyond the ability of average database software tools to capture, store, manage, and analyze them effectively. Previously, limited innovations in hardware, processing power, and data storage were the Achilles heel of Big Data analysis. Yet, recent technological innovations and advancements have helped in overcoming previous limitations to Big Data analysis.

Organizations today are collecting a broader variety of data at a greater velocity than before, resulting in immense growth in data volume. According to a 2011 study by the McKinsey Global Institute, organizations in varying sectors capture trillions of bytes of information about their customers, suppliers, and operations through digital systems. For example, millions of networked sensors embedded in mobile phones, automobiles and other products, are continually creating and communicating data. The result is a 40 percent projected annual growth in the volume of data generated. Fifteen out of 17 sectors in the U.S. economy have more data stored per company than the U.S. Library of Congress—which has collected more than 235 terabytes of data in April 2011 alone. By any definition, that's Big Data.

Clearly, we have an abundance of data. But what makes this a genuinely revolutionary moment is that thanks to Moore's Law we also have the computing power and the ability to store, retrieve, and analyze such data, all at an affordable price. Recent innovations have had the combined effect of providing commodity multi-core processors and distributed computing frameworks that allow software such as R to exploit this new generation of hardware. Other innovations that utilize these changes include external memory algorithms, data chunking, and the ability to execute these in distributed computing environments. Today's computing hardware is more than capable of crunching through datasets that were considered insurmountable just a few years back.

Advances in Analytics and Visualization through the Evolution of Second-Generation, "Big Analytics" Platforms

Until recently, only clustered samples of observations were used for statistical analysis, despite the burgeoning size and proportions of datasets. Reliance on a sample of the population within a universe of observations was a necessity for statistical analysis because of both the limitations and costs of hardware and the limitations of software to work with big data. Thus, the method of sampling compromised the degree of accuracy and depth of analysis possible.

While most organizations have invested heavily in first-generation analytics applications, recent advances in second-generation "Big Analytics" platforms such as Revolution R Enterprise have improved both analytical and organizational performance. Big Analytics platforms are optimized for Big Data and utilize today's massive datasets, thanks to recent performance advancements coupled with innovative analytic platforms. In addition to the analytic routines themselves, data visualization techniques and the way the analytics are executed on various kinds of hardware platforms have drastically improved and increased in capabilities.

Armed with the advantages of Big Data, advanced computing hardware, and a new generation of software capable of exploiting these changes, the potential impact of Big Analytics is not trivial. For example, McKinsey estimates that Big Data could help the U.S. health care industry generate $300 billion in value every year, resulting in an 8-percent reduction in national spending on health care. Big Data could help U.S. retailers boost margins by 60 percent. According to McKinsey, governments in Europe could save $149 billion by using Big Data to improve operational efficiency.

Open Source R plus Revolution Analytics: A Duo of Disruptive Forces in Analytics

Revolution R Enterprise is a driving force in the Big Analytics Revolution. It is based on "R," an open source programming language designed expressly for statistical analysis. With more than two million users and growing, R has emerged as the de facto standard for computational statistics and predictive analytics. The worldwide R community has generated more than 4,300 packages (specialized techniques and tools) that can be downloaded free from sites such as CRAN (Comprehensive R Archive Network). Basically, this means that if you are looking for an advanced analytic technique, you can probably find what you need—or something very close to it—with a few keystrokes and the click of a mouse.

Despite the many advantages of working with open-source R, one challenge is its in-memory computation model, which limits the amount of data that can be analyzed to the size of available RAM.

Building on of the unique architecture of R as a functional programming language, Revolution Analytics has engineered groundbreaking additions to R through their analytics software Revolution R Enterprise. Today, Revolution R Enterprise is the fastest, most economical, and most powerful predictive analytics product on the market.

Revolution R Enterprise also provides a highly efficient file structure along with data chunking capabilities to enable the processing of big data sets across all available computational resources, either on a single server or in a distributed environment. In addition, it includes multi-threaded external-memory algorithms for the key statistical procedures used to analyze Big Data. These two features, included in Revolution R Enterprise as the 'RevoScaleR' package, provide for analysis of virtually unlimited data sizes and an almost linear increase in performance.

Furthermore, Revolution R Enterprise provides capabilities to work with data stored in large-scale data warehousing architectures including Hadoop and IBM Netezza. Thus, inherent within the Revolution R architecture is the ability to continue integration into additional high speed, distributed computing platforms. These additions are all designed to benefit from Revolution Analytics' ongoing parallelization of statistical software to run on a wide variety of platforms.

Business and Strategic Advantages of Analyzing Big Data

The new wave in technological innovation and the rise of Big Data represents a significant turning point that provides a clear opportunity to break from past practices and revolutionize data analytics. Google's innovative MapReduce was a programming model designed to process large datasets, and served as a first step towards Big Data processing. Today, new platforms have also emerged in the market to utilize Big Data and engage in Big Analytics. What, then, can Big Analytics do that could not be accomplished with smaller datasets? Let's look at five specific advantages of moving towards Big Analytics.

Move Beyond Linear Approximation

First, Big Analytics allows one to move beyond linear approximation models towards complex models of greater sophistication. While it is true that bigger doesn't always mean better, small datasets often limit our ability to make accurate predictions and assessments. Previous limitations in technology and data limited us to certain statistical techniques, particularly in the form of linear approximation models.

Linear approximation models make causal inferences and predictions between two separate variables of interest. When the explanatory variable and the outcome variable reflect a linear relationship on an X-Y axis, linear models are an optimal tool to predict and analyze results. The method is commonly used in business environments for forecasting and financial analysis. For example, investors and traders may calculate linear models to recognize how prices, stocks, or revenues increase or decrease depending on certain factors such as timing, the level of GDP, and the like.

However, using linear models on non-linear relationships limits our ability to make accurate predictions. For example, an investor may want to determine the trend and long-term movement of GDP or stock prices over an extended period of time. However, certain events or specific periods may influence the outcome and may not be capable of being captured in a linear model. To overcome such possibilities, more sophisticated models are necessary.

Today, access to large datasets allows for more sophisticated models that are accurate and precise. An example is the manipulation of a continuous variable into a set of discrete, categorical variables. When we have a variable with a wide range of values, say annual income or corporate sales revenue, we may want to observe how specific values or intervals within the variable influence the outcome of interest. We can convert our continuous variable into discrete, categorical variables (or "factors", as we refer to them in R) by converting them into variables that takes on the value 1 if the observation is included within the category, and 0 otherwise. While increasing the complexity and sophistication of the model, the shift from a continuous to a categorical variable can better determine the exact relationship.

An example illustrates this concept: We want to determine the relationship between total years of education, total years of experience in the workforce, and annual earnings. To determine this relationship, we can rely on US Census data and obtain a sample of the population with millions of observations. Regressing years of education and years in the labor force on annual earnings as continuous variables shows that both variables have a significant and positive linear impact on annual earnings. While the linear approximation appears to work, appearances can be deceiving. In fact, the linear model above is under-predicting, inadequate and misleading. When looking at both years of education and years in the labor force, the linear approximation hides as much as it reveals.

If you convert years of education and years in the labor force into factors and regress them on annual earnings, we are able to make better predictions. It turns out that education has no discernible impact on earnings until an individual has attained a high school degree. After attaining a high school diploma, each additional year and degree (e.g., BA, MA, and professional degrees) brings higher earnings and greater rewards. This impact is reflected less accurately in a linear model. While earnings explosively grow year by year for the first 15 years in the workforce, there is very slow and marginal growth for the next decade, and income remains constant or actually declines afterwards.

Think of what such a simple change in model specification means to an organization in terms of predicting revenues, frequency of sales, or manufacturing costs and production yields. More sophisticated models can lead to greater precision and accuracy in forecasts. The reason for using traditional approximation techniques was straightforward: We had no choice and we were forced to do so because we simply didn't have enough data or the technological capabilities. However, this is no longer the case and we can vastly improve predictability when we move beyond linear approximation models with access to larger datasets, powerful technological capabilities, and most importantly, through the emergence of effective tools like Revolution R to juggle both developments. Furthermore, it's also important to note that analyzing the data set—which is large, but not large enough to qualify as Big Data – took just five seconds on a commodity 8-core laptop.

The recent turn towards Big Data and the emergence of tools like Revolution R that are capable of dealing with such data allow for greater accuracy and precision. By moving away from the limitations of past models and techniques, our ability to make predictions and assessments has drastically improved, and this advance benefits every sector of our economy. As businesses, corporations, and organizations are under greater pressure to produce more accurate and precise forecasts to maximize output, the demand of working with big data and utilizing the most efficient and effective tools have never been greater.

Data Mining and Scoring

Second, Big Analytics allows for predictive analyses utilizing techniques such as data mining and statistical modeling. When the goal of analysis is to build the best model to see the reality, the use of such techniques coupled with Big Data vastly improves predictability and scoring. These techniques can handle hundreds of thousands of coefficients based on millions of observations in large datasets to produce scores for each observation. For example, credit worthiness, purchasing propensity, trading volumes, and churn in mobile telephone clients can all be approached this way.

Here's an example, based on the data made famous in the American Statistical Association's 2009 Data Expo Challenge (http://stat-computing.org/dataexpo/2009/). The data set contains information on all U.S. domestic flights between 1987 and 2008, and includes more than 123 million observations. It nicely illustrates the opportunities that Big Data sets provide for building models that are not feasible with smaller data sets.

For example, linear regressions estimate a linear model for "time spent in the air" as a function of the interactions of origin airports and destination airports plus the effects of unique carriers and days of the week. This model, which produces over 122,000 coefficients, could not be estimated with a small dataset. If you replace "time in the air" with "minutes of daily cell phone use" or "determinates of trading volume taking into account thousands of conditions on hundreds of millions of trades," you can see the potential economic value of this approach. As more companies adopt cloud-computing strategies, predicting peak "cloud loads" will become another critical use for these kinds of newer, non-traditional analytic techniques.

Again, it is worth noting that the processing time for this problem on an 8-core standard laptop is approximately 140 seconds with Revolution Analytics' RevoScaleR package. Scoring on the basis of this model requires little or no additional time. Thus, even predictive analysis like data mining and statistical modeling is revolutionized with Big Data.

Big Data and Rare Events

Third, Big Data vastly improves the ability to locate and analyze the impact of "rare events" that might escape detection in smaller data sets with more limited variables and observations. Big Data analytics are much more likely to find the "Black Swans," or uncommon events that often get overlooked as possible outcomes.

Consider the analysis of casualty insurance claims. In a typical year, the vast majority of policyholders file no claims, and only a fraction of policyholders actually file claims. However, among this small fraction of policyholders, a smaller fraction of policyholders files costly claims of large magnitude and ultimately determines the loss or profitability of the whole portfolio. While claims related to floods, hurricanes, earthquakes, and tornadoes are relatively rare, they are also disproportionately large and costly.

Underwriters use special variants of general linear models (Poisson and Gamma regressions) to analyze claims data. These datasets are enormous in size and tax the capacity limits of legacy software by tying up powerful servers for dozens of hours to complete a single analysis. However, when using newer, predictive analytics designed with Big Data in mind, one can radically reduce the cost and time of handling these types of files. Imagine the strategic business value of being able to predict rare events with greater accuracy.

Extracting and Analyzing 'Low Incidence Populations'

Fourth, Big Data helps to extract and analyze cases that focus on low incidence populations. In many instances, data scientists have trouble locating "low incidence populations" within a large data set. Low incidence populations might be rare manufacturing failures, treatment outcomes of patients with rare diseases or extremely narrow segments of high-value customers. For example, low incidence populations can refer to individuals with genetic disorders amongst a sample population.

The challenge of locating low incidence populations is made even more difficult when only a fraction of the dataset is sampled for analysis, a prevalent concern in the past whereby data scientists utilized samples that were collected from the total population. The few observations one has to work with makes it difficult to capture the outcome of interest and make predictive assessments and forecasts. To analyze such low incidence populations requires screening and extracting numerous observations to find enough cases in what may be little more than only .001 % of a population.

In situations like these, new hardware and software systems utilizing parallel computing are capable of generating insight that go beyond the scope or ability of traditional analytics. Take 23andMe, a revolutionary personal genomics and biotechnology company at the forefront of technology utilizing the power of the internet and recent advances in DNA analysis technology to engage in genetic research. In addition to providing members with personal genetic information using web-based interactive tools, 23andMe offers members an opportunity to participate in 23andMe's research studies and advance genetic research.

Through their initiative, 23andMe has built a database of over 100,000 members, and is currently researching over 100 diseases, conditions, and traits. With their rapidly expanding genomic database and their unique ability to pinpoint specific groups of interest, 23andMe is at the forefront and intersection of scientific and technological discovery.

For example, 23andMe's most recently published study includes a discovery of the origins of two new genetic regions associated with Parkinson's disease. Furthermore, by engaging with big data and computational power, Wired Magazine found that 23andMe's research took a mere eight months to complete. In contrast, Wired estimated that a typical research project undertaken by the NIH would have taken six years to come to the same conclusion. 23andMe provides an exemplary case of how the new revolution in big data can influence scientific discovery and research.

Recent developments and the presence of Big Data allow for greater predictive and analytical power. With larger datasets, there is less concern for data scientists to make inaccurate forecasts. This has made the accuracy of extracting and analyzing low incidence populations much more effective.

Big Data and the Conundrum of Statistical Significance

Finally, Big Data allows one to move beyond inference and statistical significance and move towards meaningful and accurate analyses. With first-generation tools, analyses were based on experimentation and inference, and data scientists and statisticians sought to maximize statistical significance in their models.

The concept of statistical significance emerged out of experimental design with random assignment of small numbers of observations; necessitated by hardware constraints, limits to analytical tools, and the difficulty of obtain the entire universe of cases. Small, representative samples of the population are tested against chance and are generalized to the entire population. In this regard, the methods and tools utilized were aimed at maximizing the significance of models. Results that maximized probability estimates (or p-values) were both vital and necessary.

By contrast, the core advantage of Big Analytics is its use of massively larger datasets that represents the majority, if not most of the population of interest. Since Big Data is unbiased and incorporates the entire population, there is little need to emphasize statistical significance and allows one to move towards making accurate and meaningful analyses.

Perhaps this is the real value of Big Data—it is what it is. Big Data does not rely on and depend on techniques to "smooth" over issues, a concern that previously haunted smaller datasets that may have been biased and unreflective of the population of interest. Big Analytics implemented with second-generation tools like Revolution R Enterprise leads to more accurate, reliable, and useful analyses for decision-making in the real world.

Conclusion

As this briefing has demonstrated, Big Analytics comes with a wide range of advances and benefits and the opportunity for greater innovation in business models, products, and services. Access to larger datasets with powerful tools means greater accuracy, transparency, and predictive power. Through access to Big Analytics, data scientists and statisticians from all sectors of the economy are now empowered to experiment with the large dataset and discover new opportunities and needs, expose greater variability, and improve performance and forecasting.

The combination of increasing access to Big Data, access to affordable, high-performance hardware, and the emergence of second-generation, analytical tools like Revolution R Enterprise provides greater reasons for stakeholders to invest in Big Analytics. The convergence of these trends means that data analysts need no longer rely on traditional analytic methods that were developed largely to cope with the inherent challenges of using small data sets and running complex analyses on expensive configurations of hardware and software. More important, this convergence provides organizations with the capabilities required to analyze large data sets quickly and cost-effectively for the first time in history.

The dawn of the "Big Data Revolution" brings us to a turning point and unique moment in the history of data analysis that eclipses the traditions of the past. This revolution is neither theoretical nor trivial and represents a genuine leap forward and a clear opportunity to realize enormous gains in efficiency, productivity, revenue, and profitability. The Age of Big Analytics is here.

About the Author

Norman H. Nie is President and CEO of Revolution Analytics, a company aimed at expanding the use of R to galvanize a predictive analytics market he helped create as co-founder of SPSS. With a distinguished career in both business and academia, Nie brings strategic vision, experienced management and proven execution.

Nie co-invented the Statistical Package for Social Sciences (SPSS) while a graduate student at Stanford University and co-founded a company around it in 1967. He served as president and CEO through 1992 and as Chairman of the Board of Directors from 1992 to 2008. Under his leadership, SPSS grew to become a leader and pioneer in statistical analysis. He is also Chairman of Board of Directors for Knowledge Networks, a survey research firm he co-founded in 1997.

Nie is Professor Emeritus in Political Science at the University of Chicago and Stanford University, where he also founded the Stanford Institute for Quantitative Study of Society (SIQSS). He is a two-time recipient of the Woodrow Wilson Award (for the best book in Political Science published in the prior year) and was recognized with a lifetime achievement award by the American Association of Public Opinion Research (AAPOR). He was recently appointed Fellow of The American Academy for the Arts and Sciences. Nie was educated at the University of the Americas in Mexico City, Washington University in St. Louis and Stanford University, where he received a Ph.D. in political science.

About Revolution Analytics

Revolution Analytics delivers advanced analytics software at half the cost of existing solutions. Led by predictive analytics pioneer and SPSS co-founder Norman Nie, the company brings high performance, productivity, and enterprise readiness to open-source R, the most powerful statistics software in the world.

Leading organizations including Merck, Bank of America and Acxiom rely on Revolution R Enterprise for their data analysis, development and mission-critical production needs.

Revolution Analytics is committed to fostering the growth of the R community, and offers free licenses of Revolution R Enterprise to academia. Revolution Analytics is headquartered in Palo Alto, Calif. and backed by North Bridge Venture Partners and Intel Capital.

Contact Us

Join the R Revolution at www.RevolutionAnalytics.com
Email: info@revolutionanalytics.com
Telephone: 650-646-9545
Twitter: @RevolutionR