Implications of Big Data on Predictive Analytics

By Perry Rotella and Nigel DeFreitas

In 1965, Intel cofounder Gordon Moore observed that the number of transistors on an integrated circuit had doubled every year since the microchip was invented. Data density has doubled approximately every 18 months, and the trend is expected to continue for at least two more decades. Moore's Law now extends to the capabilities of many digital electronic devices.[1]

Year after year, we're astounded by the implications of Moore's Law — with each new version or update bringing faster and smaller computing devices. Smartphones and tablets now enable us to generate and examine significantly more content anywhere and at any time. The amount of information has grown exponentially, resulting in oversized data sets known as Big Data. Data growth has rendered traditional management tools and techniques impractical to produce meaningful results quickly. Analytics tasks that used to take minutes now take hours or time-out altogether before completing. To tame Big Data, we need new and better methods to extract actionable insights.

According to recent studies, the world's population will produce and replicate 1.8 zetabytes (or 1.8 trillion gigabytes) of data in 2011 alone — an increase of nine times the data produced five years ago. The number of files or records (such as photos, videos, and e-mail messages) is projected to grow 75 times, while the staff tasked with managing this information is projected to increase by only 1.5 times.

Organizations that adopt modern analytic methods will be better positioned to tackle problems introduced by significantly larger data sets, while enterprises that are unable to address Big Data issues successfully will struggle to achieve meaningful results.

Information once reserved for private use is now available to the public. The U.S. government has granted public access to hundreds of thousands of data sets. And the trend is global, extending to Canadian provinces and many European nations.

Every day, users of social media sites leave behind digital records in photos, videos, and comments posted online. These text messages and images contain metadata such as the date and time they were created and may even include GIS coordinates. Telematics devices provide yet another data source by recording driving behaviors and vehicular events. This data can offer better context and risk understanding and can be combined with traditional data sources to gain superior predictive insights.

However, access to more data does not equate to more insights. The insights must be extracted and analyzed to be of business use. Predictive model development methods involve running numerous iterations of the most relevant data to attain optimal results. When data sets grow to hundreds of millions of records, the time needed to perform such tasks becomes prohibitive. And the volume of data can be so large that it's impractical if not impossible to use traditional analytics platforms.

To solve the issue, a new class of massively parallel processing (MPP) systems has emerged — processing large amounts of data faster and at less cost. Users can develop and run predictive models on a single device that also hosts the data being queried, avoiding the slow process of moving data across networks. This is referred to as in-database analytics. The performance improvements offered by such analytics platforms allow for multiple iterations or tests to be conducted in a single day on very large data sets, enabling predictive model development on hundreds of millions of records.

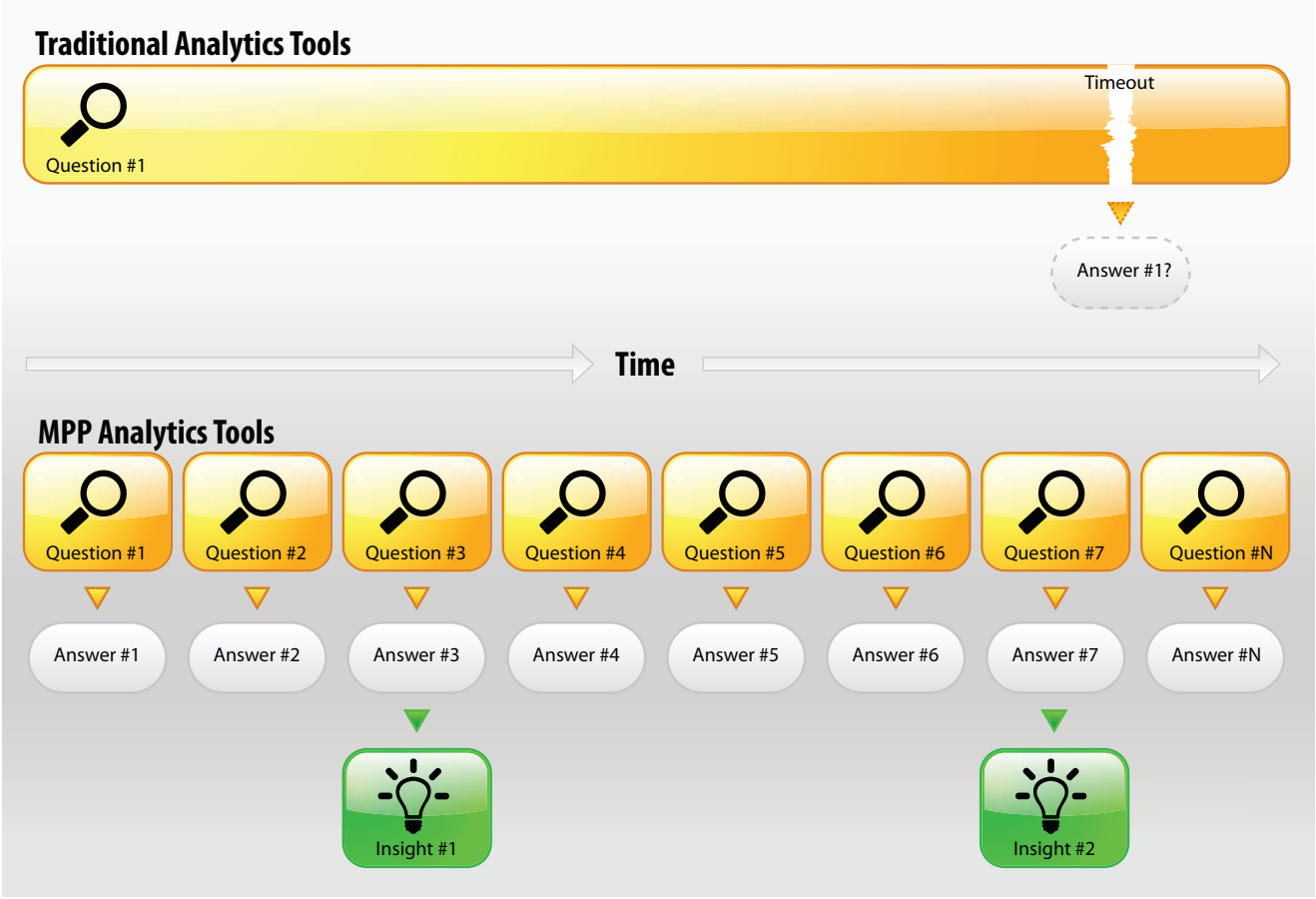
According to recent studies, the world's population will produce and replicate 1.8 zetabytes (or 1.8 trillion gigabytes) of data in 2011 alone — an increase of nine times the data produced five years ago.

When dealing with Big Data, it's important to define and follow data management principles especially as they relate to metadata, data quality, data reuse, and data security. The data manager must be aware of legal, regulatory, contractual, and business restrictions regarding the use or reuse of data. The data management experts — ideally empowered at the enterprise level — must work closely with data owners, data users, legal experts, and data security to determine what data can be combined and for what purposes. The data management group also establishes what information is available, determines where gaps exist, and assimilates newly acquired data assets into the organization.

Big Data analytics has matured to the extent that we're now able to produce answers in seconds or minutes — results that once took hours or days or were impossible to achieve using traditional analytics tools executing on older technology platforms. This ability allows modelers and business managers to gain critical insights quickly. Organizations that adopt modern analytic methods will be better positioned to tackle problems introduced by significantly larger data sets, while enterprises that are unable to address Big Data issues successfully will struggle to achieve meaningful results.

Figure 1

MPP Analytics Tools vs. Traditional Analytics Tools



Mining Big Data with traditional analytics tools is time-consuming. Complex queries time-out after hours or days, leaving critical questions unanswered. MPP analytics tools offer superior performance in tackling Big Data problems by breaking complex queries into many smaller parallel ones. MPP tools are able to return answers in seconds or minutes — leading to new discoveries and insights.

Trusted data stewards with access to centralized insurance data are uniquely situated to garner valuable insights; however, the ability to act on those insights is equally important. Organizations can distinguish themselves from the competition by developing solutions that leverage predictive models. The ability to extract more precise meaning from large data sets offers strategic and operational advantages by providing the information to make better business decisions faster.

Perry Rotella is senior vice president and chief information officer at Verisk Analytics. Nigel DeFreitas is chief application architect, strategic technology, at Verisk Analytics.

1. www.webopedia.com/TERM/M/Moores_Law.html; http://en.wikipedia.org/wiki/Moore's_law