Executive Summary

What is Big Data?

While a fog of hype often envelops the omnipresent discussions of Big Data, a clear consensus has at least coalesced around the definition of the term. "Big Data" is typically considered to be a data collection that has grown so large it can't be effectively or affordably managed (or exploited) using conventional data management tools: e.g., classic relational database management systems (RDBMS) or conventional search engines, depending on the task at hand. This can as easily occur at 1 terabyte as at 1 petabyte, though most discussions concern collections that weigh in at several terabytes at least.

Familiar Challenges, New Opportunities

If one can make one's way through the haze, it also becomes clear that Big Data is not new. Information specialists in fields like banking, telecommunications and the physical sciences have been grappling with Big Data for decades.[2] These Big Data veterans have routinely confronted data collections that outgrew the capacity of their existing systems, and in such situations their choices were always less than ideal:

- Need to access it? *Segment (silo) it.*
- Need to process it? *Buy a supercomputer.*
- Need to analyze it? *Will a sample set do?*
- Want to store it? *Forget it: use, purge, and move on.*

What is new, however, is that now new technologies have emerged that offer Big Data veterans far more palatable options, and which are enabling many organizations of all sizes and types to access and exploit Big Data for the very first time.

"In the era of big data, more isn't just more. More is different."[3]

This includes data that was too voluminous, complex or fast-moving to be of much use before, such as meter or sensor readings, event logs, Web pages, social network content, email messages and multimedia files. As a result of this evolution, the Big Data universe is beginning to yield insights that are changing the way we work and the way we play, and challenging just about everything we thought we knew about ourselves, the organizations in which we work, the markets in which we operate - even the universe in which we live.

The Internet: Home to Big Data Innovation

Not surprisingly, most of these game-changing technologies were born on the Internet, where Big Data volumes collided with a host of seemingly impossible constraints, including the need to support:

- Massive and impossible to predict traffic
- A 99.999% availability rate
- Sub-second responsiveness
- Sub-penny per-session costs
- 2-month innovation roadmaps

To satisfy these imposing requirements, Web entrepreneurs developed data management systems that achieved supercomputer power at bargain-basement cost by distributing computing tasks in parallel across large clusters of commodity servers. They also gained crucial agility – and further ramped up performance – by developing data models that were far more flexible than those of conventional RDBMS. The best known of these Web-derived technologies are non-relational databases (called "NoSQL" for "Not Only SQL," SQL being the standard language for querying

and managing RDBMS), like the Hadoop framework (inspired by Google; developed and open sourced to Apache by Yahoo!) and Cassandra (Facebook), and search engine platforms like CloudView (Exalead) and Nutch (Apache).

Another class of solutions, for which we appropriate (and expand) the "NewSQL" label coined by Matthew Aslett of the 451 Group, strives to meet Big Data needs without abandoning the core relational database model.[4] To boost performance and agility, these systems employ strategies inspired by the Internet veterans (like massive distributed scaling, in-memory processing and more flexible, NoSQL-inspired data models), or they employ strategies grown closer to (RDBMS) home, like in-memory architectures and in-database analytics. In addition, a new subset of such systems has emerged over the latter half of 2011 that goes one step further in physically combining high performance RDMBS systems with NoSQL and/or search platforms to produce integrated hardware/software appliances for deep analytics on integrated structured and unstructured data.

The Right Tool for the Right Job

Together, these diverse technologies can fulfill almost any Big Data access, analysis and storage requirement. You simply need to know which technology is best suited to which type of task, and to understand the relative advantages and disadvantages of particular solutions (usability, maturity, cost, security, etc.).

Complementary, Not Competing Tools

In most situations, NoSQL, Search and NewSQL technologies play complementary rather than competing roles. One exception is exploratory analytics, for which you may use a Search platform, a NoSQL database, or a NewSQL solution depending on your needs. A search platform alone may be all you need if 1) you want to offer self-service exploratory analytics to general business users on unstructured, structured or hybrid data, or 2) if you wish to explore previously untapped resources like log files or social media, but you prefer a low risk, cost-effective method of exploring their potential value.

Likewise, for operational reporting and analytics, you could use a Search or NewSQL platform, but Search may once again be all you need if your analytics application targets human decision-makers, and if data latency of seconds or minutes is sufficient (NoSQL systems are subject to batch-induced latency, and few situations require the nearly instantaneous, sub-millisecond latency of expensive NewSQL systems).

While a Search platform alone may be all you need for analytics in certain situations, and it is a highly compelling choice for rapidly constructing general business applications on top of Big Data, it nonetheless makes sense to deploy a search engine alongside a NoSQL or NewSQL system in every Big Data scenario, for no other technology is as effective and efficient as Search at making Big Data accessible and meaningful to human beings.

This is, in fact, the reason we have produced this paper. We aim to shed light on the use of search technology in Big Data environments – a role that's often overlooked or misunderstood even though search technologies are profoundly influencing the evolution of data management – while at the same time providing a pragmatic overview of all the tools available to meet Big Data challenges and capitalize on Big Data opportunities. Our own experience with customers and partners has shown us that for all that has been written about Big Data recently, a tremendous amount of confusion remains. We hope this paper will dispel enough of this confusion to help you get on the road to successfully exploiting your own Big Data.