Understanding Big Data: A Management Study

Special Research Reprint
Courtesy of Progress Software

SAUGATUCK TECHNOLOGY

PROGRESS SOFTWARE

TABLE OF CONTENTS

TABLE OF FIGURES

About this Report

This Big Data management report takes a broad look at how the market for Big Data infrastructure, technologies and solutions are evolving in response to the explosion in both structured and non-structure content of all types. The Cloud is emerging as a particularly useful platform for Big Data solutions for both infrastructure and analytics, as it is an ideal platform for massive data crunching and analysis at affordable prices.

Additional related research of interest is available to clients via our Research Library located at www.saugatucktechnology.com/browse-research/ (registration required). The following Saugatuck staff were instrumental in the development and publication of this report: ***Lead author:*** Brian Dooley. ***Contributing Author***: Bruce Guptill.

About Saugatuck Technology

Saugatuck Technology, Inc., provides subscription research and management consulting services focused on the key market trends and disruptive technologies driving change in enterprise IT, including Software-as-a-Service (SaaS), Cloud Infrastructure, Social Computing, Mobility and Advanced Analytics, among others. Founded in 1999, Saugatuck is headquartered in Westport, CT, with offices in Falmouth, MA, Santa Clara, CA and in Frankfurt, Germany. For more information, please visit www.saugatucktechnology.com or call +1.203.454.3900.

SUMMARY

"Big Data" is an increasingly-used but often ill-defined term, spurred in large part through the growth of Cloud IT and Cloud Business. This Saugatuck management study addresses two important aspects of Big Data for enterprise IT and business leaders:

- Determining how emerging Big Data technologies can aid them in developing real business solutions; and
- Understanding the components of Big Data solutions to make appropriate choices that meet specific requirements.

This study also enables IT providers to better understand the Big Data environment, including both the storage and access requirements and repercussions in Analytics.

SO WHAT?

Before planning for the costs and other effects of emergent trends and influences, IT and business leaders need to understand what lurks beneath the surface. To effectively and efficiently manage something, we need to understand what it's made of. How we see and manage those critical components dictate show effectively and efficiently we can manage the larger challenge. In the case of Big Data, the confluence of massive data flows, open source foundations, and Cloud IT creates a series of challenges with some built-in means of managing them – if we know what to look for and how to see it.

PERSPECTIVE

Big Data has risen to prominence recently as a result of the accelerating rate of data creation, combined with the advent of Cloud-based methods for accessing, managing, storing and analyzing extremely large volumes of data at reasonable cost. In this Saugatuck *Management Study*, we look at the growing Big Data infrastructure and analytics environments, and we examine how these components work together and fit into the emerging Big Data ecosystem. Big Data is of increasing importance as companies seek competitive advantage in the enormous data stores that are available from internal sources as well as from internet locations.

There are three areas of fundamental importance in understanding Big Data. These are:

- Big Data infrastructure
- Big Data Analytics
- Cloud facilitation of Big Data processing

While much of the current attention has been focused upon Analytics, the infrastructure issues are equally important, particularly with respect to developing a capacity to access, manage and process petabytes of data. While the basis of Analytics is Hadoop and MapReduce, the basis of infrastructure is in the database systems used to organize and store data particularly in the growing area of "NoSQL" solutions. There is considerable overlap, but the infrastructure area also include issues such as integration backup and recovery, suitability to particular types of querying, ability to handle distributed storage, and the like.

Big Data is not fundamentally a new topic; it is simply a recognition that the total volume of data residing on company servers and within accessible internet locations

is now exceeding conventional management, processing and analytic techniques.

In addition to volume, Big Data questions are also concerned with the growing importance of unstructured data, and a need for immediate results. Together, these elements are often expressed as the "three Vs": Volume, Variety, and Velocity.

As data volume, variety and velocity continue to grow, Big Data is presenting a wide range of challenges that are likely to resonate through the industry for years to come. We have previously addressed this area in our discussion of Advanced Analytics "Advanced Analytics in the Cloud: Key Issues Framing the Research Agenda" (KI-839, 28January2011) and more specifically, "Critical Characteristics of Advanced Business Analytics in the Cloud"(MKT-899, 8June2011). Big Data may also be viewed as a problem in its own right, based on the steady growth in data availability over the past several years, and the resultant struggle to process it, store it, and secure it.

Processing of very large data sets raises two fundamental and related questions:

- How can we access, store, and secure the enormous and highly differentiated data sources that are now available to companies.
- How can we process this data to derive meaningful information from it and use it in business operations.

The first question is about infrastructure, and recent debate has centered upon database structure, particularly in the NoSQL database movement. Infrastructure also includes standard questions of data storage and security, which play into the choices that need to be made. The second question involves Advanced Analytics, and how data can be effectively analyzed across extremely large data sets. This discussion has recently centered around Hadoop and MapReduce, though other analytic techniques are also of importance.

Both of these questions ultimately concern Cloud IT, which is providing the data access and storage infrastructure, the means for analysis, and a range of new possibilities which have brought attention to this area.

Data continues to burst the seams of conventional architectures and processing techniques, as digitization extends across all areas of endeavor, and companies attempt to manage, process and analyze it.

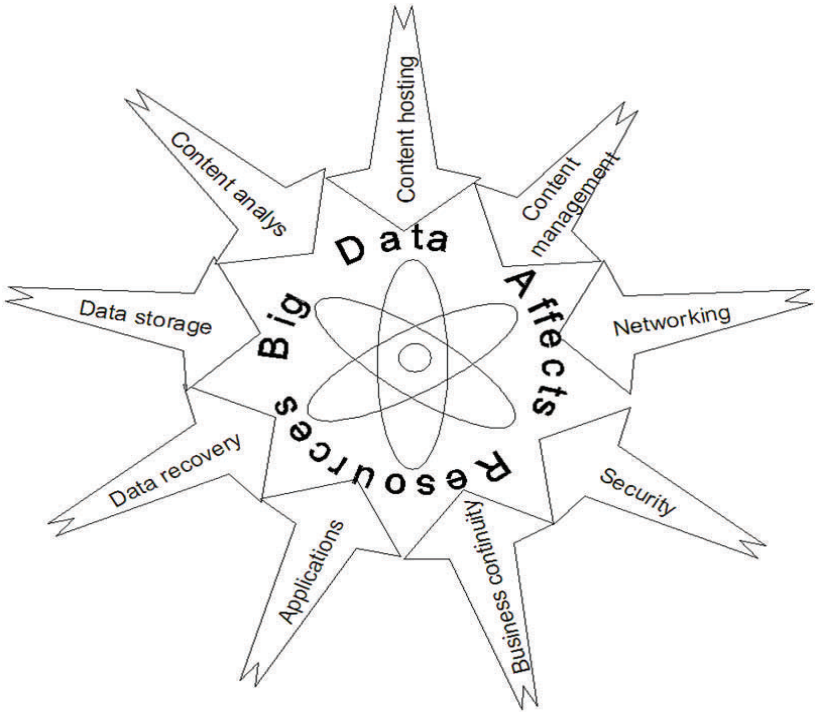
The Problem with Big Data

Big Data is about the massive growth that we have seen in digital data as everything knowable becomes digitized and new forms of communication that only exist within the digital realm continue to be added to the mix. Data has been growing very quickly for a very long time, with the conventional estimate being a doubling every 18 months. The McKinsey Global Institute has estimated that enterprises globally stored more than 7 exabytes (7×2^{60} bytes) of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks.

Volume by itself does not begin to describe the true picture, as illustrated in Figure 1. As different types of objects are brought into the data stream, such as voice, video, architectural plans, and customer comments, new issues emerge in how these items can be processed, stored and accessed, and, indeed, how they can be differentiated from each other. As we move into a world where the digital stream is largely unstructured, is not necessarily stored in an ordered way, exists in real time,

and exists in formats that have special processing requirements, the old assumptions begin to break down. This is the root of the Big Data problem.

Figure 1: Big Data Complexities



Source: Saugatuck Technology Inc.

Big Data is not just about Analytics, though this is perhaps the most urgent area. It is also about organization, categorization, and access to data. There is an increasing realization that all data is not alike, and this means that the uniform models previously used to manage, store, analyze and retrieve it in the past no longer operate so effectively. Not only is the amount much greater, but the differentiation is also greater, and techniques used to shoehorn unwilling data objects (BLOBs, for example) into unnatural arrangements soon break down when any kind of real access is required.

Extraordinary growth in data, although predictable, continues to strain corporate resources in both infrastructure and processing sectors. As Figure 1 denotes, areas affected include:

- Data storage
- Data recovery
- Applications
- Business Continuity
- Security
- Networking and network infrastructure
- Content management
- Content hosting
- Content analysis

As processing of enormous databases and multi-gigabyte artifacts becomes more common, processes themselves will advance to both provide better management and to take advantage of the rich nature of evolving digital content. Thus, data growth will continue to impact all areas.

The current areas of impact are within development of alternative database designs, particularly within the "NoSQL" movement; and within Advanced Analytics, where Hadoop and MapReduce are becoming of increasing importance and defining new market sectors.

Big Data NoSQL Databases

The traditional relational database system with SQL access was developed in an earlier era, where structured information could be accessed, categorized and normalized with relative ease. It was not designed for enormous scale, and neither was it designed for extremely rapid processing. It was designed to meet a wide array of different query types, looking at corporate data which was—and remains—processed in a highly structured way by traditional software. The idea of a record, with its fixed areas of data entry and limited information types is synonymous with this usage.

NoSQL originally stood for No SQL; today it is generally agreed that it means "not only SQL". These are database products designed to handle extremely large data sets. There are a variety of different types of database types that fall within the general NoSQL area. Perhaps the most important are the Columnar, Key/Value, and Document systems. The Columnar systems have been growing within the proprietary area, with the leading smaller players all being acquired by large database vendors.

Types of NoSQL include the following:

- Key-value systems, based on Amazon's Dynamo (2007), using a hash table with a unique key and pointer to a data item. These include Memcached, Dynamo and Voldemort. Amazon's S3 uses Dynamo as its storage mechanism.
- Columnar systems, used to store and process very large amounts of data distributed over many machines. Keys point to multiple columns. The most important example is Google's BigTable, where rows are identified by a row key with the data sorted and stored by this key. BigTable has served as the basis of a number of NoSQL systems, including Hadoop's Cassandra (open sourced from Facebook) and HBase, and Hypertable. Column based systems also include AsterData and Greenplum.
- Document Databases, based on Lotus Notes, similar to key-value, but based on versioned documents that are collections of other key-value collections. The best known of these are MongoDB and CouchDB.
- Graph Database systems, built with nodes, relationships between nodes and the properties of nodes. Instead of tables of rows and columns, a flexible graph model is used which can scale across multiple machines. An example is the open source Neo4J.

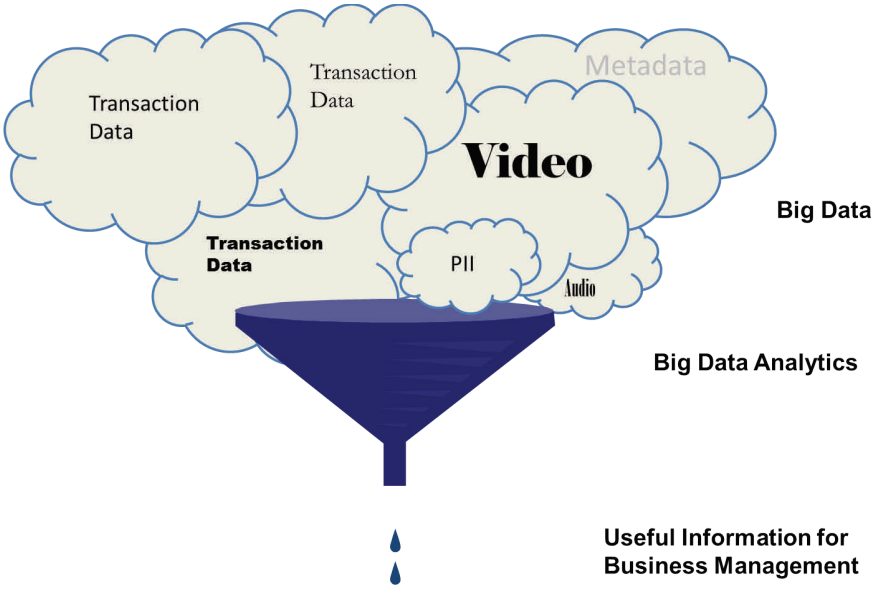
Each of these database systems has a range of advantages and disadvantages that are tied to particular types of problems and their solutions. It is instructive to note that IBM's IMS, and the CODASYL database systems, which preceded RDMS and SQL are in use today for handling very large data stores. Significantly, IBM's IMS is still used to reliably record financial transactions around the world.

One thing that has become clear is that there is no single solution to Big Data problems. Instead, there are a variety of different database models emerging that are more specialized and suitable for handling specific types of problems. For example, the columnar databases that have been popular recently are designed for high speed access to data on a distributed basis, and work well with MapReduce. But document databases, such as MongoDB and CouchDB work better with documents, and incorporate features for high speed high volume processing of document objects. Graph databases are specialized to graph data and key-value databases are another form of high speed processing format that is suitable for large data sets with relatively simple characteristics.

Big Data Analytics

Big Data Analytics is to be distinguished from general Big Data issues for a number of reasons. First, it is more about processing than about the underlying database, meaning that the discussion is more likely to be around Hadoop and MapReduce than around database types. Secondly, Big Data Analytics needs to coexist with regular analytics and Business Intelligence (BI). It is this type of concern, in fact, which led the NoSQL database movement from "No SQL" to "Not Only SQL". Big Data Analytics needs to accommodate existing analytics and Business Intelligence, and integrate with data from these sources. It is being added to the major RDMS and Analytics solutions by IBM, HP, Microsoft, Oracle, PeopleSoft and so forth.

Figure 2: Big Data, Analytics, and Useful Business Information



Source: Saugatuck Technology Inc.

While Big Data is being driven by the sheer volume of data that companies need to organize and store, Big Data Analytics is driven by the desire to understand that data and develop usable information from it. Since much of the data is unstructured, that means that problems of processing depth are added to volume. For example in sentiment analysis of Social Networking, there may be billions of records, and each made up of natural language which must be individually dissected for meaning. So processing of each

record presents challenges even before the issues of aggregation of results might be considered.

The pre-eminent tool for Big Data Analytics has been Hadoop, based on MapReduce. This open source platform has been incorporated in a range of open source and proprietary analytic products This has a number of advantages over other forms of processing, including open source availability, standardization, usability over a fairly wide range of problems, recent evolution, and suitability to current IT infrastructure. However, it is important to bear in mind that the problems associated with Big Data Analytics are not necessarily new; they are simply more commonplace, and more urgent. Many of the issues have been seen within the HPC and Grid Computing areas in the past.

Hadoop Market

The size of the Hadoop market in itself distinguishes this sector from other processing methods for Big Data Analytics. In recent years, it has become a central focus for discussion, and it has spawned an ecosystem that now includes both open source and proprietary solutions, as well as methods for emulation and integration.

Hadoop provides non-SQL high performance processing in a multiprocessor-efficient system for handling complex queries. Its parallel programming model hides the complexities of distribution and fault tolerance; programming is eased by availability of a number of utilities such as Hive and Pig from the Hadoop project, plus a variety of tools from external sources. Key components of Hadoop are its MapReduce processing component, which provides parallel processing capability for enormous data sets; and the Hadoop Distributed File System (HDFS), which apportions the task to processing nodes.

The Apache Hadoop project includes a number of related open source Apache projects that include Pig, Hive, Cassandra, HBase, Avro, Chukwa, Mahout and Zookeeper. Of these projects, Hive and Pig are most familiar, as they are frequently used in Hadoop projects. The NoSQL databases HBase and Cassandra are used as the database grounding for a significant number of Hadoop projects.

As Hadoop has risen in prominence as a Big Data architecture, competing types of processes also need to be mentioned, many coming out of decades of work in the HPC and Grid Computing territories. This is particularly important where Big Data meets the Cloud, as discussed in "Cloud IT Effects on Advanced Analytics" (MKT-885, 5May2011).

Of particular importance in considering the role of Hadoop/MapReduce in analytics is the fact that this type of processing is inherently batch-like, and not immediately suitable for real time analysis. It is also unsuited to ad-hoc queries. Hadoop solves the Volume issue of the three Vs, but it needs help to solve Velocity (real time processing) and Variety (differing object types).

Integration Issues

The key to Big Data Analytics is integration with other Analytic and BI solutions, This means that there is an ongoing effort to accommodate SQL, as well as to add the strengths of the data warehouse to the Big Data Solution. Accommodation has ranged from creation of SQL-Alike or SQL-Extended query languages to use of Hadoop to extract data for insertion into data warehouses as a "super ETL" utility. Recent mergers and acquisitions have highlighted this strategy, most notably with Aster Data being acquired by Terradata. and rival Greenplum acquired by EMC.

Vendors have adopted numerous strategies for accommodating both SQL and Hadoop, including embedding SQL in MapReduce applications (Greenplum), adding traditional capabilities on top of Hadoop (IBM, Pentaho, Jaspersoft), proving a Hadoop connector for RDMS (Aster Data), and layering SQL on top of Hadoop (Hadoop Hive). This strategy makes it possible. Including MapReduce in analytic RDMS platforms potentially offers some of the best of both worlds.

Another piece of the puzzle comes into play with specially adapted hardware used to create Big Data processing appliances. The importance of this approach has been indicated most strongly by IBM with its acquisition of appliance vendor Netezza and development of the Watson Q&A system that was used to win at Jeopardy, displaying prowess at big data processing, rapidity of response, and natural language parsing. These systems tend to use Hadoop, but almost as an afterthought. In effect, Big Data Analytics appliances are HPC-in-a-cloud devices that have specialized to perform a range of analytic tasks, with processing efficiency inbuilt at the hardware level. HP has also been operating in this area, along with a number of smaller specialty firms.

Hadoop Alternatives

Hadoop has gained much recent attention due to its availability, history of use, and applicability to some of the key problems in large scale analytics. However, it is not the only solution, and neither is it the first. Problems involving Big Data have been around for a long time, particularly in scientific computing, and many solutions have been found for specific problem types within areas such as High Performance Computing (HPC) and Grid Computing.

Hadoop works best with a specific range of processing tasks that are mainly fairly simple and do not involve complex joins, ACID requirements, or real time access. Hadoop and its ecosystem have been developing to meet these challenges, and numerous utilities and Hadoop variations exist that address these issues. However, it is important to note that the RDBMS-based data warehousing environment has also been developing to meet the challenges of Big Data analytics, including various methods of incorporating Hadoop and MapReduce, plus specialty database systems such as VoltDB. From the HPC space, MPI and BSP provide parallel programming capabilities for complex algorithms, and have been used for many years in solving Big Data problems. New capabilities being deployed and made available as open source by online companies - which have an urgent requirement for Big Data analysis - include Google's Dremel, Pregel and Percolator.

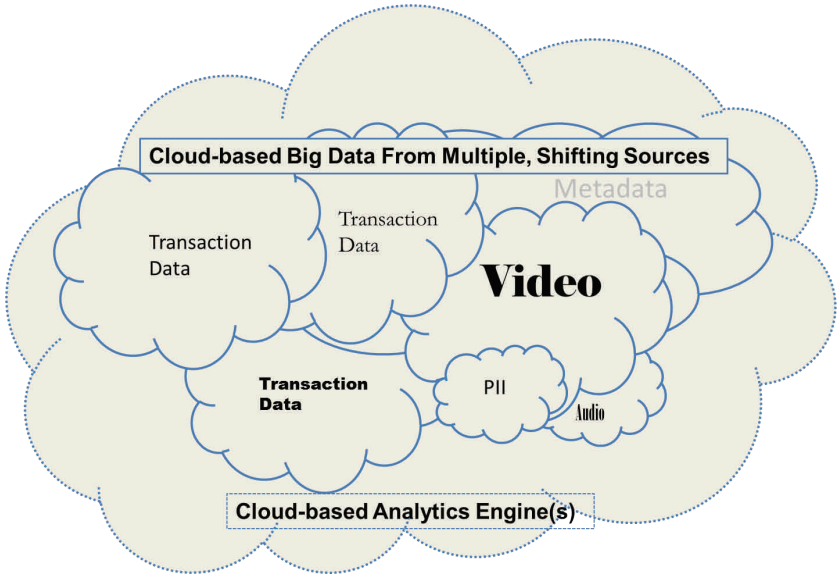
Big Data and the Cloud

The Cloud has emerged as a principal facilitator of Big Data, both at the infrastructure and at the analytic levels. As we have previously described, the Cloud offers a range of options for Big Data Analysis in both public and private cloud settings. On the infrastructure side, Cloud IT provides options for managing and accessing very large data sets as well as for supporting powerful infrastructure elements at relatively low cost.

The Cloud is particularly well suited to Big Data operations. The virtual, amorphous nature of Cloud IT – adaptable, flexible, and powerful – certainly lends itself to the enormous and shifting environment(s) of Big Data, as seen in Figure 3. Cloud architectures consist of arrays of virtual machines that are ideal for the processing of very

large data sets, to the extent that processing can be segmented into numerous parallel processes. This affinity was discovered at an early stage of Cloud IT development, frequently leading directly to development of Hadoop clusters that could be used for analytics.

Figure 3: Cloud-Based Analytics Can Envelope, Adapt to, and Contain Big Data



Source: Saugatuck Technology Inc.

Many of the commercial Big Data problems involve online data such as click-streams for advertising and consumer comments for marketing, making them particularly suitable to processing through Cloud-based solutions. A wide variety of these solutions now exist, as discussed in "Cloud IT Effects on Advanced Analytics" (MKT-885, 5May2011). Hadoop, for example, is offered directly from the Cloud by Cloudera and Amazon. Amazon's Elastic MapReduce offers a hosted Hadoop framework on its IaaS and PaaS offerings, and open source vendor Cloudera offers its Cloudera Distribution for Hadoop (CDH) over Amazon Web Services. Also, the major BI and Analytics vendors continue to expand their cloud-based Advanced Analytics solutions, including Big Data processing and Hadoop support, across private clouds, public clouds, and hybrid clouds. IBM, Microsoft, Oracle and HP have all been highly active in this sector.

Cloud IT is likely to become increasingly important as an enabler of Big Data, both for storage and access and for analytics. Development of Hybrid Clouds capable of integrating public data and private corporate data is particularly critical. Most Big Data applications will depend upon the capability to bring together external and corporate data to provide usable information. Additionally, processing of multi-petabyte data stores will be highly dependent on local storage capability, with processing in-situ rather than requiring large scale data movement and ETL. The Cloud therefore provides a point of access as well as a mechanism for integration between private corporate data warehouses and processing of public data. Its virtualized architecture enables the parallel processing needed for solving these problems, and there will be an increasing number of SaaS solutions capable of performing the processing and data integration tasks.

Current Trends

Big Data is a growing problem for corporations. Current discussion has frequently centered around Hadoop and NoSQL, but these are only a starting point. There is a growing realization that there is no single solution to the problems that are being faced in this area. Instead, there is a range of available solutions that need to be tuned to the specific problems being addressed. Hadoop has clearly emerged as an important component of Big Data solutions, but it is not a panacea, and there is widespread divergence in how it is being used, how the results are being integrated with other data, and how programming and usage are facilitated. There are also a range of other processing methods that should be examined with respect to the problem at hand.

NoSQL databases have also arisen as a component of the Big Data problem, particularly under the argument that data and processing requirements now exceed the capabilities of standard RDBMS and Analytic RDBMS systems. While this is to some extent true, it is more significant to look at the wide range of current alternative data stores, which include columnar databases, IMS and CODASYL solutions, and the range of NoSQL variations. Integration with current corporate data will be critical, and this means that even these solutions must eventually be brought into the corporate data warehouse and analytics solution. The major RDMS suppliers are now all providing some accommodation. Structured corporate data will not go away, and neither will it lose its usefulness; the synergies between Big Data solutions and current analytics solutions will provide a key area of growth for tomorrow.

In addition to integration issues, Big Data problem are likely to involve an increasing array of new problems in data location, security, processing and purpose that have only begun to emerge. For example, storage of Personally Identifiable Data/ Personally Identifiable Information (PID/PII) is illegal in many areas, but what about data and processing that permits such data to be generated? Companies are now only at the very edge of this emerging area, as the vast wave of digital data enables new approaches, new information, and generates new territories of risk.

Net Impact

The problems of (and with) Big Data are likely to increase as a result of sheer data volume along with radical changes in the types of data being stored and analyzed, and its characteristics. IT organizations are moving into new territory, where the relatively simple tabular RDMS tools of the past are no longer sufficient to meet processing requirements. The situation is similar to when RDMS systems and then Data Warehousing were developed as solutions to data access problems that no longer suited CODASYL.

New strategies for handling and analyzing data are now emerging fro handling Big Data problems. New solutions need to provide adequate storage and access and easily distribute workloads across multiple machines. They also need to integrate a wide variety of different data types. RDBMSes are excellent for the purposes that they have served and will continue to serve—handling corporate structured data and serving it up to a wide variety of applications and ad-hoc access requirements. However, it is clear that new forms of data, particularly those involving enormous quantities of data and unstructured items, also need to be stored and analyzed.

Ultimately, both SQL and NoSQL need to come together, and Hadoop needs to be accommodated along with SQL. This is already beginning to happen, with the major

database vendors acquiring and integrating a variety of Advanced Analytic solutions, including NoSQL. A variety of ways of integrating this data and creating new analytic schemes is already beginning to appear.

For the user, Big Data represents a wide variety of opportunities for exploring customer data, predicting market directions, understanding patterns in location data, and developing predictive models. Integration of Big Data processing is also placing strains upon storage, networks and infrastructure that need to be better understood.

For Vendors, Big Data provides a snapshot of the growing infrastructure needs of emerging data management and analytics solutions. It points to the areas that need to be strengthened and the solutions that need to be put into place to meet the next generation of integrated Data Management an Analytics solutions.

The problems and opportunities that result as data increases in volume, variety, and velocity are only just beginning to emerge. Solutions are developing across the IT environment. The current focus has been upon database development and basic processing, but developments here are shaping the infrastructure of the future and enabling a wide variety of new questions that may be asked of stored and accessible data.

SAUGATUCK OFFERINGS AND SERVICES

Saugatuck Technology provides subscription research / advisory and consulting services to senior business and IT executives, technology and software vendors, business / IT services providers, and investors.

Our Mission is to help our clients make better business decisions and create new business value through trusted and objective insights into the key market trends and emerging technologies driving real change.

Over the last few years, this has included a major focus on Software-as-a-Service (SaaS), Cloud Infrastructure, and Social Computing, among other key trends.

CONTINUOUS RESEARCH SERVICES (CRS)

- Subscription research / advisory services that provide independent / unbiased analysis, insights and guidance into the most important emerging technologies driving change in business computing.
- We are experts in *Cloud Business* and *Cloud IT,* among other key market trends / technologies - with a balanced view that is valued by both providers and consumers of technology-enabled products / services.

USER STRATEGIC CONSULTING SERVICES

- Leadership and Planning Workshops
- Strategy and Program Assessments
- Vendor Selection / Evaluations
- Cloud Transition / Migration and Mgmt Best Practices

VENDOR STRATEGIC CONSULTING SERVICES

- Market Assessment
- Strategy Validation
- Opportunity Analysis
- Positioning / Messaging / Go-to-Market Strategies
- Competitive Analysis

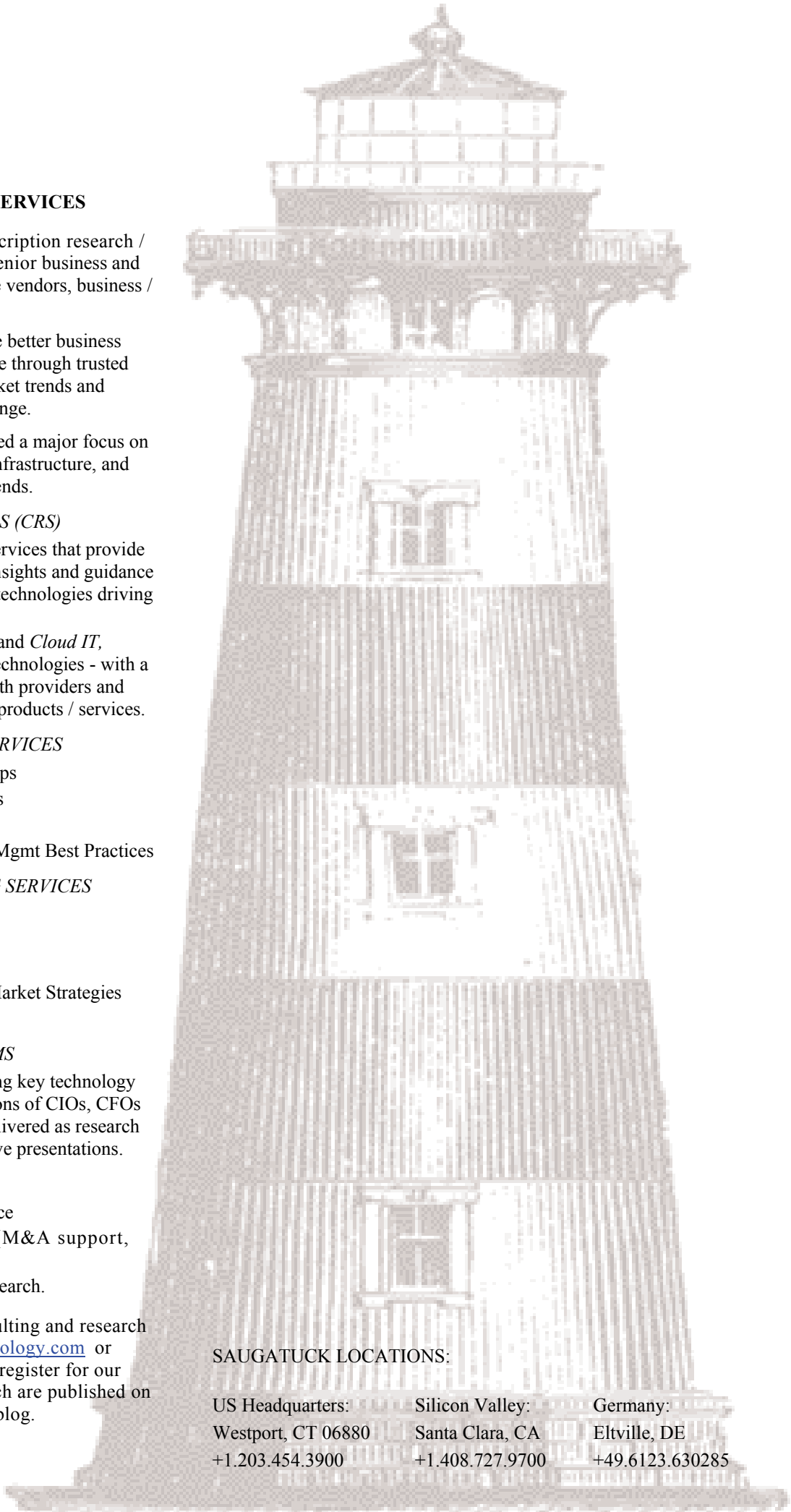
THOUGHT-LEADERSHIP PROGRAMS

- Custom research programs targeting key technology and business/IT investment decisions of CIOs, CFOs and senior business executives, delivered as research reports, position papers or executive presentations.

VALUE-ADDED SERVICES

- Competitive and market intelligence
- Investment advisory services (M&A support, due diligence)
- Primary and Secondary market research.

To learn more about Saugatuck consulting and research offerings, go to www.saugatucktechnology.com or email Chris MacGregor. While there register for our complimentary *Research Alerts*, which are published on a weekly basis, or visit our Lens360 blog.



SAUGATUCK LOCATIONS:

US Headquarters:	Silicon Valley:	Germany:
Westport, CT 06880	Santa Clara, CA	Eltville, DE
+1.203.454.3900	+1.408.727.9700	+49.6123.630285