DASSAULT SYSTEMES

EXALEAD

A PRACTICAL GUIDE TO BIG DATA

Opportunities, Challenges & Tools

"Give me a lever long enough and a fulcrum on which to place it, and I shall move the world." [1]

Archimedes

3DS.COM/EXALEAD

ABOUT THE AUTHOR

Laura Wilber is the former founder and CEO of California-based AVENCOM, Inc., a software development company specializing in online databases and database-driven Internet applications (acquired by Red Door Interactive in 2004), and she served as VP of Marketing for Kintera, Inc., a provider of SaaS software to the nonprofit and government sectors. She also developed courtroom tutorials for technology-related intellectual property litigation for Legal Arts Multimedia, LLC. Ms. Wilber earned an M.A. from the University of Maryland, where she was also a candidate in the PhD program, before joining the federal systems engineering division of Bell Atlantic (now Verizon) in Washington, DC. Ms. Wilber currently works as solutions analyst at EXALEAD. Prior to joining EXALEAD, Ms. Wilber taught Business Process Reengineering, Management of Information Systems and E-Commerce at ISG (l'Institut Supérieur de Gestion) in Paris. She and her EXALEAD colleague Gregory Grefenstette recently co-authored *Search-Based Applications: At the Confluence of Search and Database Technologies*, published in 2011 by Morgan & Claypool Publishers.

ABOUT EXALEAD

Founded in 2000 by search engine pioneers, EXALEAD® is the leading Search-Based Application platform provider to business and government. EXALEAD's worldwide client base includes leading companies such as PricewaterhouseCooper, ViaMichelin, GEFCO, the World Bank and Sanofi Aventis R&D, and more than 100 million unique users a month use EXALEAD's technology for search and information access. Today, EXALEAD is reshaping the digital content landscape with its platform, EXALEAD CloudView™, which uses advanced semantic technologies to bring structure, meaning and accessibility to previously unused or under-used data in the new hybrid enterprise and Web information cloud. CloudView collects data from virtually any source, in any format, and transforms it into structured, pervasive, contextualized building blocks of business information that can be directly searched and queried, or used as the foundation for a new breed of lean, innovative information access applications.

EXALEAD was acquired by Dassault Systèmes in June 2010. EXALEAD has offices in Paris, San Francisco, Glasgow, London, Amsterdam, Milan and Frankfurt.

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 constraints, 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 applicances 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 instanteous, 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.

TABLE OF CONTENTS

1) CROSSING THE ZETTA FRONTIER

Fueled by the pervasiveness of the Internet, unprecedented computing power, ubiquitous sensors and meters, addictive consumer gadgets, inexpensive storage and (to-date) highly elastic network capacity, we humans and our machines are cranking out digital information at a mind-boggling rate.

IDC estimates that in 2010 alone we generated enough digital information worldwide to fill a stack of DVDs reaching from the earth to the moon and back. That's about 1.2 zettabytes, or more than one trillion gigabytes—a 50% increase over 2009.[5] IDC further estimates that from 2011 on, the amount of data produced globally will double every 2 years.

No wonder then scientists coined a special term – "Big Data" – to convey the extraordinary scale of the data collections now being amassed inside public and private organizations and out on the Web.

A. What Exactly is "Big Data"?

Big Data is more a concept than a precise term. Some apply the "Big Data" label only to petabyte-scale data collections (> one million GB). For others, a Big Data collection may house 'only' a few dozen terabytes of data. More often, however, Big Data is defined situationally rather than by size. Specifically, a data collection is considered "Big Data" when it is so large an organization cannot effectively or affordably manage or exploit it using conventional data management tools.

BIG DATA
A data collection that is too large to be effectively or affordably managed using conventional technologies.

Measuring Big Data
Disk Storage*

1000 Gigabytes (GB)	≈	1 Terabyte (To)
1000 Terabytes	≈	1 Petabyte (Po)
1000 Petabytes	≈	1 Exabyte (Eo)
1000 Exabytes	≈	1 Zettabyte (Zo)
1000 Zettabytes	≈	1 Yottabyte (Yo)

* For Processor or Virtual Storage, replace 1000 with 1024.

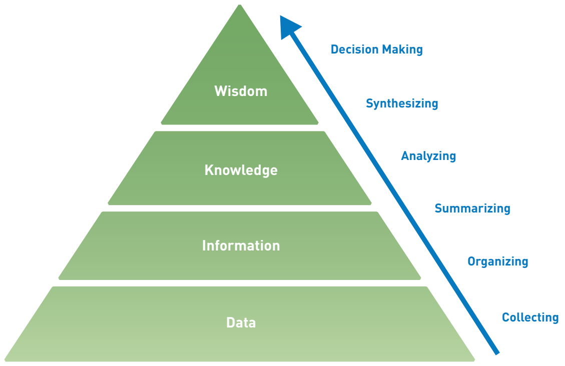
B. Who Is Affected By Big Data?

Big Data has been of concern to organizations working in select fields for some time, such as the physical sciences (meteorology, physics), life sciences (genomics, biomedical research), government (defense, treasury), finance and banking (transaction processing, trade analytics), communications (call records, network traffic data), and, of course, the Internet (search engine indexation, social networks).

Now, however, due to our digital fecundity, Big Data is becoming an issue for organizations of all sizes and types.
In fact, in 2008 businesses were already managing on average 100TB or more of digital content.[6] Big Data has even become a concern of individuals as awareness grows of the depth and breadth of personal information being amassed in Big Data collections (in contrast, some, like LifeLoggers,[7] broadcast their day-to-day lives in a Big Data stream of their own making).

C. Big Data: Boon or Bane?

For some, Big Data simply means Big Headaches, raising difficult issues of information system cost, scaling and performance, as well as data security, privacy and ownership.
But Big Data also carries the potential for breakthrough insights and innovation in business, science, medicine and government—if we can bring humans, machines and data together to reveal the natural information intelligence locked inside our mountains of Big Data.



The classic data management mission: transforming raw data into action-guiding wisdom. In the era of Big Data, the challenge is to find automated, industrial-grade methods for accomplishing this transformation.

2) BIG DATA OPPORTUNITIES

Innovative public and private organizations are already demonstrating that transforming raw Big Data collections into actionable wisdom is possible. They are showing in particular that tremendous value can be extracted from the "grey" data that makes up the bulk of Big Data, that is to say data that is unused (or under-used) because it has historically been:
1) Too voluminous, unstructured and/or raw (i.e., minimally structured) to be exploited by conventional information systems, or
2) In the case of highly structured data, too costly or complex to integrate and exploit (e.g., trying to gather and align data from dozens of databases worldwide).

These organizations are also opening new frontiers in operational and exploratory analytics using structured data (like database content), semi-structured data (such as log files or XML files) and unstructured content (like text documents or Web pages).

Some of the specific Big Data opportunities they are capitalizing on include:

- Faceted search at scale
- Multimedia search
- Sentiment analysis
- Automatic database enrichment
- New types of exploratory analytics
- Improved operational reporting

We'll now look more closely at these opportunities, with each accompanied by a brief example of an opportunity realized using a technology whose role is often overlooked or misunderstood in the context of Big Data: the search engine. We'll then review the full range of tools available to organizations seeking to exploit Big Data, followed by further examples from the search world.

A. Faceted Search at Scale

Faceted search is the process of iteratively refining a search request by selecting (or excluding) clusters or categories of results. In contrast to the conventional method of paging through simple lists of results, faceted search (also referred to as parametric search and faceted navigation) offers a remarkably effective means of searching and navigating large volumes of information—especially when combined with user aids like type-ahead query suggestions, auto-spelling correction and fuzzy matching (matching via synonyms, phonetics and approximate spelling).

NATURAL LANGUAGE PROCESSING (NLP)

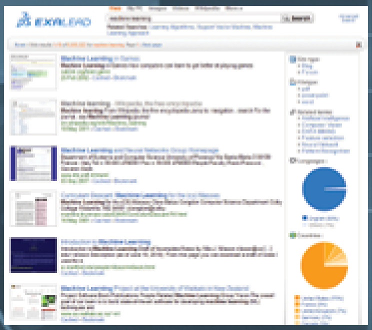
Rooted in artificial intelligence, NLP—also referred to as computational linguistics—uses tools like statistical algorithms and machine learning to enable computers to understand instances of human language (like speech transcripts, text documents and SMS messages). While NLP focuses on the structural features of an utterance, semantics goes beyond form in seeking to identify and understand meanings and relationships.

Until recently, faceted search could only be provided against relatively small data sets because the data classification and descriptive meta-tagging upon which faceted search depends were largely manual processes. Now, however, industrial-grade natural language processing (NLP) technologies are making it possible to automatically classify and categorize even Big Data-size collections of unstructured content, and hence to achieve faceted search at scale.

FACETED SEARCH EXAMPLE:

EXALEAD CloudView™ uses industrial-grade semantic and statistical processing to automatically cluster and categorize search results for an index of 16 billion Web pages (approx. 6 petabytes of raw data).

Facets hide the scale and complexity of Big Data collections from end users, boosting search success and making search and navigation feel simple and natural.



You can see industrial faceting at work in the dual Web/enterprise search engine EXALEAD CloudView^{TM}, in other public Web search engines like Google, Yahoo! and Bing, and, to varying degrees of automation and scale, in search utilities from organizations like HP, Oracle, Microsoft and Apache.

Look for this trend to accelerate and to bring new accessibility to unstructured Big Data.

B. Multimedia Search

Multimedia content is the fastest growing type of user-generated content, with millions of photos, audio files and videos uploaded to the Web and enterprise servers daily. Exploiting this type of content at Big Data scale is impossible if we must rely solely on human tagging or basic associated metadata like file names to access and understand content.

However, recent technologies like automatic speech-to-text transcription and object-recognition processing (called Content-Based Image Retrieval, or CBIR) are enabling us to structure this content from the inside out, and paving the way toward new accessibility for large-volume multimedia collections. Look for this trend to have a significant impact in fields like medicine, media, publishing, environmental science, forensics and digital asset management.

Multimedia Search Example:
FRANCE 24 is a 24/7 international news channel broadcasting in French, English and Arabic. In partnership with EXALEAD, Yacast Media and Vecsys, FRANCE 24 is automatically generating near real-time transcripts of its broadcasts, and using semantic indexation of these transcripts to offer "full text" search inside videos. Complementary digital segmentation technology enables users to jump to the precise point in the broadcast where their search term is used.

C. Sentiment Analysis

Sentiment analysis uses semantic technologies to automatically discover, extract and summarize the emotions and attitudes expressed in unstructured content. Semantic analysis is sometimes applied to behind-the-firewall content like email messages, call recordings and customer/constituent surveys. More commonly, however, it is applied to the Web, the world's first and foremost Big Data collection and the most comprehensive repository of public sentiment concerning everything from ideas and issues to people, products and companies.

The Web: The world's first and foremost Big Data collection.

Sentiment analysis on the Web typically entails collecting data from select Web sources (industry sites, the media, blogs, forums, social networks, etc.), cross-referencing this content with target entities represented in internal systems (services, products, people, programs, etc.), and extracting and summarizing the sentiments expressed in this cross-referenced content.

SENTIMENT ANALYSIS EXAMPLE:
A large automotive vehicle manufacturer uses Web sentiment analysis to improve product quality management. The application uses the EXALEAD CloudView^{TM} platform to extract, analyze and organize pertinent quality-related information from consumer car forums and other resources so the company can detect and respond to potential issues at an early stage. Semantic processors automatically structure this data by model, make, year, type of symptom and more.

This type of Big Data analysis can be a tremendous aid in domains as diverse as product development and public policy, bringing unprecedented scope, accuracy and timeliness to efforts such as:

- Monitoring and managing public perception of an issue, brand, organization, etc. (called "reputation monitoring")
- Analyzing reception of a new or revamped service or product
- Anticipating and responding to potential quality, pricing or compliance issues
- Identifying nascent market growth opportunities and trends in customer demand

D. Database Enrichment

Once you can collect, analyze and organize unstructured Big Data, you can use it to enhance and contextualize existing structured data resources like databases and data warehouses. For instance, you can use information extracted from high-volume sources like email, chat, website logs and social networks to enrich customer profiles in a Customer Relationship Management (CRM) system. Or, you can extend a digital product catalog with Web content (like, product descriptions, photos, specifications, and supplier information). You can even use such content to improve the quality of your organization's master data management, using the Web to verify details or fill in missing attributes.

DATABASE ENRICHMENT EXAMPLE:
The travel and tourism arm of France's high speed passenger rail service, Voyages-SNCF, uses unstructured Web data (like local events and attractions and travel articles and news) to enhance the content in its internal transport and accommodation databases. The result is a full-featured travel planning site that keeps the user engaged through each stage of the purchase cycle, boosting average sales through cross-selling, and helping to make Voyages-SNCF.com a first-stop reference for travel planning in France.

E. Exploratory Analytics

Exploratory analytics has aptly been defined as "the process of analyzing data to learn about what you don't know to ask."[8] It is a type of analytics that requires an open mind and a healthy sense of curiosity. In practice, the analyst and the data engage in a two-way conversation, with researchers making discoveries and uncovering possibilities as they follow their curiosity from one intriguing fact to another (hence the reason exploratory analytics are also called "iterative analytics").

In short, it is the opposite of conventional analytics, referred to as Online Analytical Processing (OLAP). In classic OLAP, one seeks to retrieve answers to precise, pre-formulated questions from an orderly, well-known universe of data. Classic OLAP is also sometimes referred to as Confirmatory Data Analysis (CDA) as it is typically used to confirm or refute hypotheses.

> *"Big Data will become a key basis of competition, underpinning new waves of productivity growth, innovation and consumer surplus—as long as the right policies and enablers are in place."*
>
> McKinsey Global Institute [9]

Discovering Hidden Meanings & Relationships

There is no doubt that the Big Data collections we are now amassing hold the answers to questions we haven't yet thought to ask. Just imagine the revelations lurking in the 100 petabytes of climate data at the DKRZ (German Climate Computing Center), or in the 15 petabytes of data produced annually by the Large Hadron Collider (LHC) particle accelerator, or in the 200 petabytes of data Yahoo! has stocked across its farm of 43,000 (soon to be 60,000) servers.

> *"[Exploratory analytic] techniques make it feasible to look for more haystacks, rather than just the needle in one haystack."* [10]

An even richer vein lies in cross-referencing individual collections. For example, cross-referencing Big Data collections of genomic, demographic, chemical and biomedical information might move us closer to a cure for cancer. At a more mundane level, such large scale cross-referencing may simply help us better manage inventories, as when Wal-Mart hooked up weather and sales data and discovered that hurricane warnings trigger runs not just on flashlights and batteries (expected), but also on strawberry Pop-Tarts breakfast pastries (not expected), and that the top-selling pre-hurricane item is beer (surprise again).

However, Wal-Mart's revelation was actually not the result of exploratory analytics (as is often reported), but rather conventional analytics. In 2004, with Hurricane Frances on the way, Wal-Mart execs simply retrieved sales data for the days before the recently passed Hurricane Charley from their then-460TB data warehouse, and fresh supplies of beer and pastries were soon on their way to stores in Frances' path.[11]

What's important about the Wal-Mart example is to imagine what could happen if we could turn machines loose to discover such correlations on their own. In fact, we do this now in two ways: one can be characterized as a "pull" approach, the other a "push" strategy.

In the "pull" method, we can turn semantic mining tools loose to identify the embedded relationships, patterns and meanings in data, and then use visualization tools, facets (dynamic clusters and categories) and natural language queries to explore these connections in a completely ad hoc manner. In the second "push" method, we can sequentially ask the data for answers to specific questions, or instruct it to perform certain operations (like sorting), to see what turns up.

Improving the Accuracy and Timeliness of Predictions

The goal of exploratory, "let's see what turns up" analytics is almost always to generate accurate, actionable predictions. In traditional OLAP, this is done by applying complex statistical models to clean sample data sets within a formal, scientific "hypothesize, model, test" process.

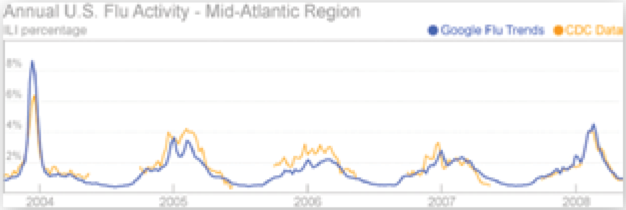
> *"Business decisions will increasingly be made, or at least corroborated, on the basis of computer algorithms rather than individual hunches."* [12]

Exploratory analytics accelerate this formal process by delivering a rich mine of ready-to-test models that may never have otherwise come to light. And, though conventional predictive analytics are in no danger of being sidelined, running simple algorithms against messy Big Data collections can produce forecasts that are as accurate as complex analytics on well-scrubbed, statistically-groomed sample data sets.

For example, real estate services provider Akerys uses the EXALEAD CloudView^{TM} platform to aggregate, organize and structure real estate market statistics extracted daily from the major real estate classifieds websites. As a result, Akerys' public Labo-Immo project (labo-immo.org) enables individuals to accurately identify and explore market trends two-to-three months in advance of the official statistics compiled by notaries and other industry professionals.

Search-based analytics offers an effective means of distilling information intelligence from large-volume data sets, especially un- or semi-structured corpora such as Web collections.

In another example drawn from the world of the Web, Google analyzed the frequency of billions of flu symptom-related Web searches and demonstrated that it was possible to predict flu outbreaks with as much accuracy as the U.S. Centers for Disease Control and Prevention (CDC), whose predictions were based on a complex analytics applied to data painstakingly compiled from clinics and physicians. Moreover, as people tend to conduct Internet research before visiting a doctor, the Web search data revealed trends earlier, giving health care communities valuable lead time in preparing for outbreaks. Now the CDC and other health organizations like the World Health Organization use Google Flu Trends as an additional disease monitoring tool.[13]



Trends as an additional disease monitoring tool.

Of note, too, is the fact that neither the CDC nor clinic directors care why Web searches so closely mirror—and anticipate—CDC predictions: they're just happy to have the information. This is the potential of exploratory Big Data analytics: sample it all in, see what shows up, and, depending on your situation, either act on it—or relay it to specialists for investigation or validation.

> *"Invariably, simple models and a lot of data trump more elaborate models based on less data."* [14]
>
> Alon Halevy, Peter Norvig & Fernando Pereira

F. Operational Analytics

While exploratory analytics are terrific for planning, operational analytics are ideal for action.The goal of such analytics is to deliver actionable intelligence on meaningful operational metrics in real or near-real time.

This is not easy as many such metrics are embedded in massive streams of small-packet data produced by networked devices like 'smart' utility meters, RFID readers, barcode scanners, website activity monitors and GPS tracking units. It is machine data designed for use by other machines, not humans.

Making it accessible to human beings has traditionally not been technically or economically feasible for many organizations. New technologies, however, are enabling organizations to overcome technical and financial hurdles to deliver human-friendly and analysis of real-time Big Data streams (see Chapter 4).

As a result, more organizations (particularly in sectors like telecommunications, logistics, transport, retailing and manufacturing) are producing real-time operational reporting and analytics based on such data, and significantly improving agility, operational visibility, and day-to-day decision making as a result.

Consider, for example, the case of Dr. Carolyn McGregor of the University of Ontario. Conducting research in Canada, Australia and China, she is using real-time, operational analytics on Big Data for early detection of potentially fatal infections in premature babies. The analytics platform monitors real-time streams of data like respiration, heart rate and blood pressure readings captured by medical equipment (with electrocardiograms alone generating 1,000 readings per second).

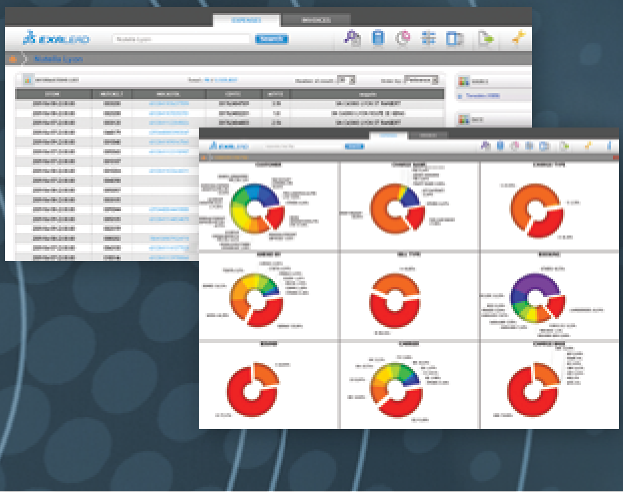
The system can detect anomalies that may signal the onset of an infection long before symptoms emerge, and well in advance of the legacy approach of having a doctor review limited data sets on paper every hour or two. As Dr. McGregor notes, *"You can't see it with the naked eye, but a computer can."* [15]

EXPLORATORY ANALYTICS EXAMPLE:

In an example of exploratory analytics inside the enterprise, one of the world's largest global retailers is using an EXALEAD CloudView^{TM} Search-Based Application (SBA) to enable non-experts to use natural language search, faceted navigation and visualization to explore the details of millions of daily cash register receipts. Previously, these receipts, which are stored in an 18TB Teradata data warehouse, could only be analyzed by Business Intelligence system users executing canned queries or complex custom queries.

A second SBA further enables users to perform exploratory analytics on a cross-referenced view of receipt details and loyalty program data (also housed in a Teradata data warehouse). Users can either enter a natural language query like "nutella and paris" to launch their investigations, or they can simply drill down on the dynamic data clusters and categories mined from source systems to explore potentially significant correlations.

Both of these SBAs are enabling a wide base of business users to mine previously siloed data for meaningful information. They are also improving the timeliness and accuracy of predictions by revealing hidden relationships and trends.



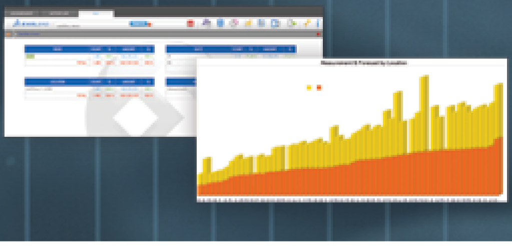
OPERATIONAL ANALYTICS EXAMPLE:
A leading private electric utility and the world's largest renewable energy operator has deployed a CloudView Search-Based Application (SBA) to better manage its wind power production. Specifically, they are using CloudView to automate cumbersome analytic processes and deliver timelier production forecasts.

The CloudView SBA works by allowing a quasi-real-time comparison of actual production data from metering equipment (fed into an Oracle system) and forecast data produced by an MS SQL Server application. Prior to deploying CloudView, separately stored production and forecast data had to be manually compared – an inefficient and error-prone process with undesirable lag time.

The new streaming predictive analytics capability is boosting the company's ability to achieve an optimal balance between actual and forecast production to minimize costly surpluses or deficits. The use of an SBA also offers unlimited, ad-hoc drill down on all data facets maintained in source systems, including reporting and analytics by geographic location (country, region, city, etc.) and time period (hour, day, week, month, etc.). Historical data is retained for long-range analytics.

As an added benefit, the platform is improving overall information systems responsiveness by offloading routine information requests from the Oracle and MS SQL Server systems. The Proof-of-Concept (POC) for this SBA was developed in just 5 days.

See the GEFCO and La Poste case studies in Chapter 5 for additional examples of operational reporting and analytics on Big Data.



3) BREAKTHROUGH INNOVATION FROM THE INTERNET

As the examples in Chapter 2 demonstrate, it is possible to overcome the technical and financial challenges inherent in seizing Big Data opportunities. This capability is due in large part to tools and technologies forged over the past 15 years by Internet innovators including:

- Web search engines like EXALEAD, Google and Yahoo!, who have taken on the job of making the ultimate Big Data collection, the Internet, accessible to all.
- Social networking sites like LinkedIn and Facebook.
- eCommerce giants like Amazon.

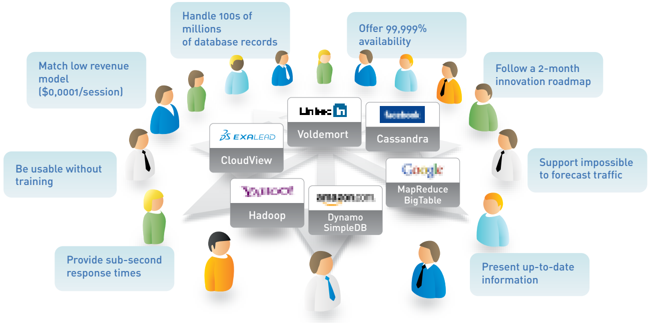
> *"Reliability at massive scale is one of the biggest challenges we face at Amazon.com... Even the slightest outage has significant financial consequences and impacts customer trust."* [16]
>
> Amazon

These organizations and others like them found that conventional relational database technology was too rigid and/or costly for many data processing, access and storage tasks in the highly fluid, high-volume world of the Web.

Relational database management systems (RDBMS) were, after all, initially designed (half a century ago) to accurately and reliably record transactions like payments and orders for brick-and-mortar businesses. To protect the accuracy and security of this information, they made sure incoming data adhered to elaborate, carefully-constructed data models and specifications through processing safeguards referred to as ACID constraints (for data Atomicity, Consistency, Isolation and Durability).

These ACID constraints proved to be highly effective at ensuring data accuracy and security, but they are very difficult to scale, and for certain types of data interaction—like social networking, search and exploratory analytics—they are not even wholly necessary. Sometimes, maximizing system availability and performance are higher priorities than ensuring full data consistency and integrity.

Accordingly, Internet businesses developed new data management systems that relaxed ACID constraints and permitted them to scale their operations massively and cost-effectively while maintaining optimal availability and performance.


Handle 100s of millions of database records
Offer 99,999% availability
Match low revenue model ($0,0001/session)
Follow a 2-month innovation roadmap
Voldemort
Cassandra
CloudView
Be usable without training
Support impossible to forecast traffic
Hadoop
Dynamo SimpleDB
MapReduce BigTable
Provide sub-second response times
Present up-to-date information

Internet Drives Data Management Innovation

SOME COMMON RDBMS

MS SQL Server
MySQL
PostgreSQL
Oracle 11g
IBM DB2 & Informix

ACID CONSTRAINTS

Atomicity
Consistency
Isolation
Durability

A. Distributed Architectures & Parallel Processing

One of the most important ways they achieved this was by distributing processing and access tasks in parallel across large (and often geographically dispersed) grids of loosely coupled, inexpensive commodity servers.

Working in parallel, these collections of low-end servers can rival supercomputers in processing power at a fraction of the cost, and ensure continuous service availability in the case of inevitable hardware failures.

It is an architecture inspired by symmetric multi-processing (SMP), massively parallel processing (MPP) and grid computing strategies and technologies.

B. Relaxed Consistency & Flexible Data Models

In addition to distributed architectures and parallel processing, these Internet innovators also achieved greater performance, availability and agility by designing systems that can ingest and process inconsistent, constantly evolving data. These flexible models, together with semantic technologies, have also played a primary role in making grey data exploitable (these models are discussed in Chapter 4, Section B, Data Processing & Interaction).

TYPES OF PARALLEL PROCESSING

In parallel processing, programming tasks are broken into subtasks and executed in parallel across multiple computer processors to boost computing power and performance. Parallel processing can take place in a single multiple processor computer, or across thousands of single- or multi-processor machines.

SMP is parallel processing across a small number of tightly-coupled processors (e.g., shared memory, data bus, disk storage (sometimes), operating system (OS) instance, etc.).

MPP is parallel processing across a large number of loosely-coupled processors (each node having its own local memory, disk storage, OS copy, etc.). It is a "shared nothing" versus "shared memory" or "shared disk" architecture. MMP nodes usually communicate across a specialized, dedicated network, and they are usually homogeneous machines housed in a single location.

Grid Computing also employs loosely-coupled nodes in a shared-nothing framework, but, unlike SMP and MPP, a grid is not architected to act as a single computer but rather to function like individual collaborators working together to solve a single problem, like modeling a protein or refining a climate model.

Grids are typically inter-organizational collaborations that pool resources to create a shared computing infrastructure. They are usually heterogeneous, widely dispersed, and communicate using standard WAN technologies. Examples include on-demand grids (e.g., Amazon EC2), peer-to-peer grids (e.g., SETI@Home), and research grids (e.g., DutchGrid).

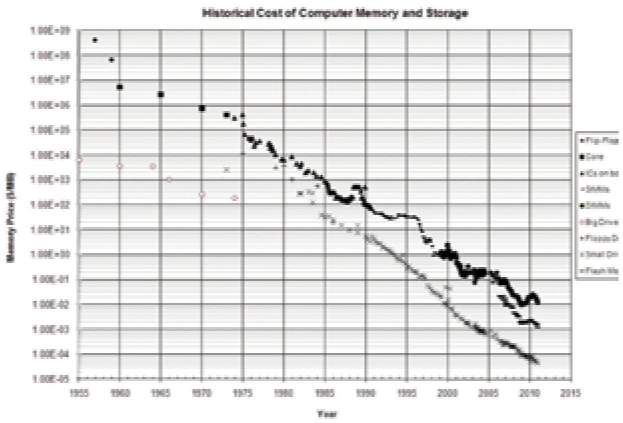
C. Caching & In-Memory Processing

Most further developed systems that make heavy use of data caching, if not full in-memory storage and processing. (In in-memory architectures, data is stored and processed in high speed RAM, eliminating the back-and-forth disk input/output (I/O) activity that can bottleneck performance.) This evolution is due in equal parts to innovation, a dramatic decrease in the cost of RAM (see chart below), and to the rise of distributed architectures (even though the price of RAM has dropped, it's still far less expensive to buy a batch of commodity computers whose combined RAM is 1TB than to buy a single computer with 1TB RAM).

While few organizations deals with Internet-scale data management challenges, these Web-born innovations have nonetheless spawned pragmatic commercial and open source tools and technologies anyone can use right now to address Big Data challenges and take advantage of Big Data opportunities.

Let's look at that toolbox now.

Graph of Memory Prices Decreasing with Time (1957-2010)



Data/chart copyright 2001, 2010, John C. McCallum.
See www.jcmit.com/mem2010.htm.

4) THE BIG DATA TOOLBOX

While some research organizations may rely on supercomputers to meet their Big Data needs, our toolbox is stocked with tools accessible to organizations of all sizes and types.
These tools include:

A.Data Capture & Preprocessing
1.ETL (Extract, Transform and Load) Tools
2.APIs (Application Programming Interfaces) / Connectors
3.Crawlers
4.Messaging Systems

B.Data Processing & Interaction
1.NoSQL Systems
2.NewSQL Systems
3.Search Engines

C.Auxiliary Tools
1.Cloud Services
2.Visualization Tools

Each has a different role to play in capturing, processing, accessing or analyzing Big Data. Let's look first at data capture and preprocessing tools.

A. Data Capture & Preprocessing

1.ETL TOOLS

Primary Uses

- Data consolidation (particularly loading data warehouses)
- Data preprocessing/normalization

Definition

ETL (Extract, Transform and Load) tools are used to map and move large volumes of data from one system to another. They are most frequently used as data integration aids. More specifically, they are commonly used to consolidate data from multiple databases into a central data warehouse through bulk data transfers. ETL platforms usually include mechanisms for "normalizing" source data before transferring it, that is to say, for performing at least the minimal processing needed to align incoming data with the target system's data model and specifications, and removing duplicate or anomalous data.

Examples

Solutions range from open source platforms to expensive commercial offerings, with some ETLs available as embedded modules in BI and database systems. Higher-end commercial solutions are most likely to offer features useful in Big Data contexts, like data pipelining and partitioning, and compatibility with SMP, MPP and grid environments.

Some ETL examples include:

- Ab Initio
- CloverETL (open source)
- IBM Infosphere DataStage
- Informatica PowerCenter
- Jasper ETL (open source – Talend-powered)
- MS SQL Server Integration Services
- Oracle Warehouse Builder (embedded in Oracle 11g) & Oracle Data Integrator
- Talend Open Studio (open source)

Caveats

In Big Data environments, the Extract process can sometimes place an unacceptable burden on source systems, and the Transform stage can be a bottleneck if the data is minimally structured or very raw (most ETL platforms require an external or add-on module to handle unstructured data). The Load process can also be quite slow even when the code is optimized for large volumes. This is why ETL transfers, which are widely used to feed data warehouses, tend to be executed during off-hours—usually overnight—resulting in unacceptable data latency in some situations. Note, however, that many ETL vendors are developing - or have already developed - special editions to address these limitations, such as the Real Time Edition of Informatica's PowerCenter (in fact, their new 9.1 release is specially tailored for Big Data environments).

2.APIS

Primary Use

- Data exchange/integration

Definition

An Application Programming Interface (API) is a software-to-software interface for exchanging almost every type of service or data you can conceive, though we focus here on the use of APIs as tools for data exchange or consolidation. In this context, an API may enable a host system to receive (ingest) data from other systems (a "push" API), or enable others to extract data from it (a publishing or "pull" API). APIs typically employ standard programming languages and protocols to facilitate exchanges (e.g., HTTP/REST, Java, XML). Specific instances of packaged APIs on a system are often referred to as "connectors," and may be general in nature, like the Java Database Connectivity (JDBC) API for connecting to most common RDBMS, or vendor/platform specific, like a connector for IBM Lotus Notes.

Examples

APIs are available for most large websites, like Amazon, Google (e.g., AdSense, Maps), Facebook, Flickr, Twitter, and MySpace. They are also available for most enterprise business applications and data management systems. Enterprise search engines usually offer packaged connectors encompassing most common file types and enterprise systems (e.g., XML repositories, file servers, directories, messaging platforms, and content and document management systems).

Caveats

With Big Data loads, APIs can cause bottlenecks due to poor design or insufficient computing or network resources, but they've generally proven to be flexible and capable tools for exchanging large-volume data and services. In fact, you could argue the proliferation of public and private APIs has played an important role in creating today's Big Data world.

Nonetheless, you can still sometimes achieve better performance with an embedded ETL tool than an API, or, in the case of streaming data, with a messaging architecture (see Messaging Systems below).

Moreover, APIs are generally not the best choice for collecting data from the Web. A crawler is a better tool for that task (see Crawlers below). There are three main drawbacks to APIs in the Web context:

- In spite of their proliferation, only a tiny percentage of online data sources are currently accessible via an API.
- APIs usually offer access to only a limited portion of a site's data.
- Formats and access methods are at the owner's discretion, and may change at any time. Because of this variability and changeability, it can take a significant amount of time to establish and maintain individual API links, an effort that can become completely unmanageable in Big Data environments.

3.CRAWLERS

Primary Use

- Collection of unstructured data (often Web content) or small packet data

Definition

A crawler is a software program that connects to a data source, methodically extracts the metadata and content it contains, and sends the extracted content back to a host system for indexation.

One type of crawler is a file system crawler. This kind of crawler works its way recursively through computer directories, subdirectories and files to gather file content and metadata (like file path, name, size, and last modified date). File system crawlers are used to collect unstructured content like text documents, semi-structured content like logs, and structured content like XML files.

Another type of crawler is a Web (HTTP/HTTPS) crawler. This type of crawler accesses a website, captures and transmits the page content it contains along with available metadata (page titles, content labels, etc.), then follows links (or a set visitation list) to proceed to the next site.

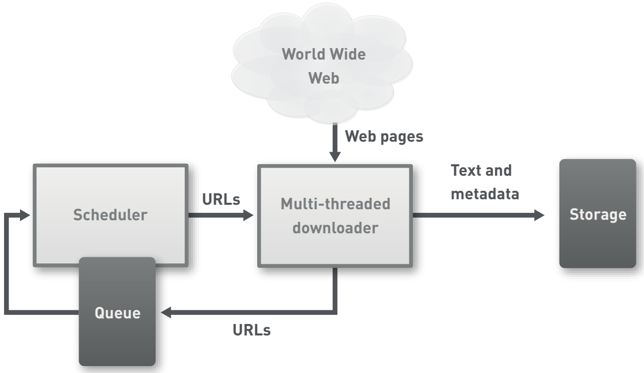
Typically a search engine is used to process, store and access the content captured by crawlers, but crawlers can be used with other types of data management systems (DMS).

Examples

File system crawlers are normally embedded in other software programs (search engines, operating systems, databases, etc.). However, there are a few available in standalone form: RiverGlass EssentialScanner, Sonar, Methabot (these are also Web crawlers).

Web crawlers are likewise usually embedded, most often in search engines, though there are standalone open source crawlers available as well. The best-known Web crawlers are those employed by public WWW search engines. Web crawler examples include:

- Bingbot
- crawler4j
- EXALEAD Crawler
- Googlebot
- Heritrix
- Nutch
- WebCrawler
- Yahoo! Slurp



Basic architecture of a standard Web crawler. Source: Wikipedia.

Caveats

As with other data collection tools, one needs to configure crawls so as not to place an undue load on the source system – or the crawler. The quality of the crawler determines the extent to which loads can be properly managed.

It should also be kept in mind that crawlers recognize only a limited number of document formats (e.g., HTML, XML, text, PDF, etc.).If you want to use a crawler to gather non-supported document formats, you'll need to convert data into an ingestible format using tools like API connectors (standard with most commercial search engines), source-system export tools, ETL platforms or messaging systems.

You should also be aware of some special challenges associated with Web crawling:

- Missed Content

Valuable data on the Web exists in unstructured, semi-structured and structured form, including Deep Web content that is dynamically generated as a result of form input and/or database querying. Not all engines are capable of accessing this data and capturing its full semantic logic.

- Low Quality Content

While crawlers are designed to cast a wide net, with backend search engines (or other DMS) being responsible for separating the wheat from the chaff, overall quality can nevertheless be improved if a crawler can be configured to do some preliminary qualitative filtering, for example, excluding certain document types, treating the content of a site as a single page to avoid crowding out other relevant sources (website collapsing), detecting and applying special rules for duplicate and near duplicate content, etc.

- Performance Problems

Load management is especially important in Web crawling. If you don't (or can't) properly regulate the breadth and depth of a crawl according to your business needs and resources, you can easily encounter performance problems. You can likewise encounter performance issues if you don't (or can't) employ a refined update strategy, zeroing in on pertinent new or modified content rather than re-crawling and re-indexing all content.

Of course, regardless of the size of the crawl, you also should avoid placing an undue load on the visited site or violating data ownership and privacy policies. These infractions are usually inadvertent and due to weaknesses in the crawler used, but they can nonetheless result in your crawler being blocked, or "blacklisted," from public websites. For internal intranet crawls, such poor management can cause performance and security problems.

In the case of the public Web, an RSS ("Really Simple Syndication" or "Rich Site Summary") feed for delivering authorized, regularly changing Web content may be available to help you avoid some of these pitfalls. But they are not available for all sites, and they may be incomplete or out of date.

MAKING SENSE OF THE WEB

A search engine sometimes views HTML content as an XML tree, with HTML tags as branches and text as nodes, and uses rules written in the standard XML query language, XPath, to extract and structure content. This is a strategy in which the crawler plays an important role in pre-processing content. A search platform may also view HTML as pure text, relying on semantic processing within the core of the engine to give the content structure.

The first approach can produce high quality results, but it is labor-intensive, requiring specific rules to be drafted and monitored for each source (in the fast-changing world of the Web, an XPath rule has an average lifespan of only 3 months). The second approach can be applied globally to all sites, but it is complex and error prone. An ideal strategy balances the two, exploiting the patterns of structure that do exist while relying on semantics to verify and enrich these patterns.

4.MESSAGING SYSTEMS

Primary Uses

- Data exchange (often event-driven, small-packet data)
- Application/systems integration
- Data preprocessing and normalization (secondary role)

Definition

Message-Oriented Middleware (MOM) systems provide an enabling backbone for enterprise application integration. Often deployed within service-oriented architectures (SOA), MOM solutions loosely couple systems and applications through a bridge known as a message bus. Messages (data packets) managed by the bus may be configured for point-to-point delivery (message queue messaging) or broadcast to multiple subscribers (publish-subscribe messaging). They vary in their level of support for message security, integrity and durability.

Exchanges between disparate systems are possible because all connected systems (“peers”) share a common message schema, set of command messages and infrastructure (often dedicated). Data from source systems is transformed to the degree necessary to enable other systems to consume it, for example, binary values may need to be converted to their textual (ASCII) equivalents, or session IDs and IP addresses may be extracted from log files and encoded as XML records. APIs for managing this data processing may be embedded in individual systems connected to the bus, or they may be embedded in the MOM platform.

Complex Event Processing (CEP)

MOM systems are often used to manage the asynchronous exchange of event-driven, small-packet data (like barcode scans, stock quotes, weather data, session logs and meter readings) between diverse systems. In some instances, a Complex Event Processing (CEP) engine may be deployed to analyze this data in real time, applying complex trend detection, pattern matching and causality modeling to streaming information and taking action as prescribed by business rules. For instance, a CEP engine may apply complex algorithms to streaming data like ATM withdrawals and credit card purchases to detect and report suspicious activity in real time or near real time. If a CEP offers historical processing, data must be captured and stored in a DMS.

Examples

MOM platforms may be standalone applications or they may be bundled within broader SOA suites. Examples include:

- Apache ActiveMQ
- Oracle/BEA MessageQ
- IBM WebSphere MQ Series
- Informatica Ultra Messaging
- Microsoft Message Queuing (MSMQ)
- Solace Messaging & Content Routers
- SonicMQ from Progress Software
- Sun Open Message Queue (OpenMQ)
- Tervela Data Fabric HW & SW Appliances
- TIBCO Enterprise Message Service & Messaging Appliance

Most of the organizations above also offer a CEP engine. There are also a number of specialty CEP vendors, including StreamBase Systems, Aleri-Coral8 (now under the Sybase umbrella), UC4 Software and EsperTech. In addition, many of the NewSQL platforms discussed in the next section incorporate CEP technology, creating uncertainty as to whether CEP will continue as a standalone technology.

Caveats

Messaging systems were specifically designed to meet the high-volume, high-velocity data needs of industries like finance, banking and telecommunications. Big Data volumes can nonetheless overload some MOM systems, particularly if the MOM is performing extensive data processing—filtering, aggregation, transformation, etc—at the message bus level. In such situations, performance can be improved by offloading processing tasks to either source or destination systems. You could also upgrade to an extreme performance solution like IBM

WebSphere MQ Low Latency Messaging or Informatica Ultra Messaging, or to a hardware-optimized MOM solution like the Solace, Tervela or TIBCO messaging appliances (TIBCO's appliance was developed in partnership with Solace).

B. Data Processing & Interaction

Today, classic RDBMS are complemented by a rich set of alternative DMS specifically designed to handle the volume, variety, velocity and variability of Big Data collections (the so-called "4Vs" of Big Data). These DMS include NoSQL, NewSQL and Search-based systems. All can ingest data supplied by any of the capture and preprocessing tools discussed in the last section (ETLs, APIs, crawlers or messaging systems).

- NoSQL

NoSQL systems are distributed, non-relational databases designed for large-scale data storage and for massively-parallel data crunching across a large number of commodity servers. They can support multiple activities, including exploratory and predictive analytics, ETL-style data transformation, and non-mission-critical OLTP (for example, managing long-duration or inter-organization transactions). Their primary drawbacks are their unfamiliarity, and, for the youngest of these largely open-source solutions, their instability.

- NewSQL

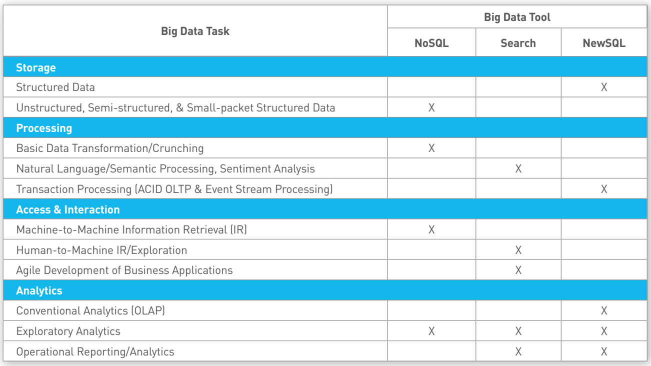
NewSQL systems are relational databases designed to provide ACID-compliant, real-time OLTP and conventional SQL-based OLAP in Big Data environments. These systems break through conventional RDBMS performance limits by employing NoSQL-style features such as column-oriented data storage and distributed architectures, or by employing technologies like in-memory processing, SMP or MPP (some go further and integrate NoSQL or Search components to address the 4V challenges of Big Data). Their primary drawback is cost and rigidity (most are integrated hardware/software appliances).

- Search-Based Platforms

As they share the same Internet roots, Big Data-capable search platforms naturally employ many of the same strategies and technologies as their NoSQL counterparts (distributed architectures, flexible data models, caching, etc.) – in fact, some would argue they are NoSQL solutions, but this classification would obscure their prime differentiator: natural language processing (NLP). It is NLP technology that enables search platforms to automatically collect, analyze, classify and correlate diverse collections of structured, unstructured and semi-structured data.

NLP and semantic technologies also enable Search platforms to do what other systems cannot: sentiment analysis, machine learning, unsupervised text analysis, etc. Search platforms are deployed as a complement to NoSQL and NewSQL systems, giving users of any skill level a familiar, simple way to search, analyze or explore the Big Data collections they house. In some situations, Search-Based Applications (SBAs) even offer an easier, more affordable alternative to NoSQL and NewSQL deployments.

As noted in the Executive Summary, the challenge with these technologies is determining which is best suited to a particular type of task, and to understand the relative advantages and disadvantages of particular solutions (usability, maturity, cost, security, technical skills required, etc.). Based on such considerations, the chart below summarizes general best use (not only use!) scenarios.



Big Data Task	Big Data Tool		
	NoSQL	Search	NewSQL
Storage			
Structured Data			X
Unstructured, Semi-structured, & Small-packet Structured Data	X		
Processing			
Basic Data Transformation/Crunching	X		
Natural Language/Semantic Processing, Sentiment Analysis		X	
Transaction Processing (ACID OLTP & Event Stream Processing)			X
Access & Interaction			
Machine-to-Machine Information Retrieval (IR)	X		
Human-to-Machine IR/Exploration		X	
Agile Development of Business Applications		X	
Analytics			
Conventional Analytics (OLAP)			X
Exploratory Analytics	X	X	X
Operational Reporting/Analytics		X	X

For the two categories with multiple options checked–exploratory and operational analytics–the choice of NoSQL, Search or NewSQL depends on whether your target user is 1) a machine, or 2) a human, and if it is a human being, whether that user is a business user or an expert analyst, statistician or programmer. The second factor is whether batch-processing or streaming analytics are right for your needs, and if streaming, whether your latency requirements are real-time, quasi-real-time or simply right-time.

To learn more, let's look more closely now at these three types of DMS.

1.NOSQL SYSTEMS

Primary Uses

- Large-scale data processing (parallel processing over distributed systems)
- Embedded IR (basic machine-to-machine information look-up & retrieval)
- Exploratory analytics on semi-structured data (expert level)
- Large volume data storage (unstructured, semi-structured, small-packet structured)

Definition

NoSQL, for "Not Only SQL," refers to an eclectic and increasingly familiar group of non-relational data management systems (e.g., Hadoop, Cassandra and BerkeleyDB). Common features include distributed architectures with parallel processing across large numbers of commodity servers, flexible data models that can accommodate inconsistent/changeable data, and the use of caching and/or in-memory strategies to boost performance. They also use non-SQL languages and mechanisms to interact with data (though some now feature APIs that convert SQL queries to the system's native query language or tool).

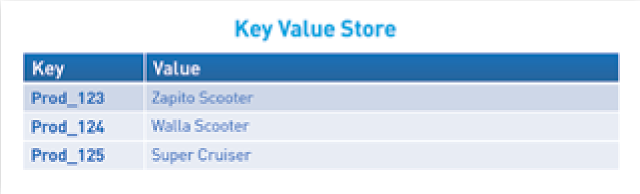
Accordingly, they provide relatively inexpensive, highly scalable storage for high-volume, small-packet historical data like logs, call-data records, meter readings, and ticker snapshots (i.e., "big bit bucket" storage), and for unwieldy semi-structured or unstructured data (email archives, xml files, documents, etc.). Their distributed framework also makes them ideal for massive batch data processing (aggregating, filtering, sorting, algorithmic crunching (statistical or programmatic), etc.).They are good as well for machine-to-machine data retrieval and exchange, and for processing high-volume transactions, as long as ACID constraints can be relaxed, or at least enforced at the application level rather than within the DMS.

Finally, these systems are very good exploratory analytics against semi-structured or hybrid data, though to tease out intelligence, the researcher usually must be a skilled statistician working in tandem with a skilled programmer.

If you want to deploy such a system in a standalone version on commodity hardware, and you want to be able to run full-text searches or ad-hoc queries against it, or to build business applications on top of it, or in general simply make the data it contains accessible to business users, then you need to deploy a search engine along with it.

NoSQL DMS come in four basic flavors, each suited to different kinds of tasks: [17]

- Key-Value stores
- Document databases (or stores)
- Wide-Column (or Column-Family) stores
- Graph databases



Key Value Store

Key	Value
Prod_123	Zapito Scooter
Prod_124	Walla Scooter
Prod_125	Super Cruiser

Most Key-Value Stores pair simple string keys with string values for fast information retrieval.

Key-Value Stores

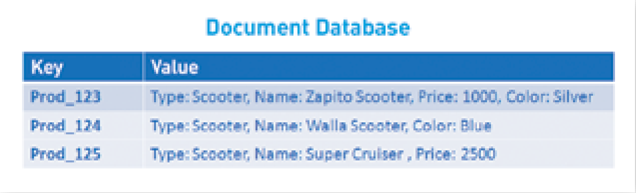
Typically, these DMS store items as alpha-numeric identifiers (keys) and associated values in simple, standalone tables (referred to as "hash tables"). The values may be simple text strings or more complex lists and sets. Data searches can usually only be performed against keys, not values, and are limited to exact matches.

Primary Use

The simplicity of Key-Value Stores makes them ideally suited to lightning-fast, highly-scalable retrieval of the values needed for application tasks like managing user profiles or sessions or retrieving product names. This is why Amazon makes extensive use of its own K-V system, Dynamo, in its shopping cart.

Examples: Key-Value Stores

- Dynamo (Amazon)
- Voldemort (LinkedIn)
- Redis
- BerkeleyDB
- Riak
- MemcacheDB



Document Database

Key	Value
Prod_123	Type: Scooter, Name: Zapito Scooter, Price: 1000, Color: Silver
Prod_124	Type: Scooter, Name: Walla Scooter, Color: Blue
Prod_125	Type: Scooter, Name: Super Cruiser , Price: 2500

Document Databases contain semi-structured values that can be queried. The number and type of attributes per row can vary, offering greater flexibility than the relational data model.

Document Databases

Inspired by Lotus Notes, document databases were, as their name implies, designed to manage and store documents. These documents are encoded in a standard data exchange format such as XML, JSON (Javascript Option Notation) or BSON (Binary JSON). Unlike the simple key-value stores described above, the value column in document databases contains semi-structured data – specifically attribute name/value pairs. A single column can house hundreds of such attributes, and the number and type of attributes recorded can vary from row to row. Also, unlike simple key-value stores, both keys and values are fully searchable in document databases.

Primary Use

Document databases are good for storing and managing Big Data-size collections of literal documents, like text documents, email messages, and XML documents, as well as conceptual "documents" like de-normalized (aggregate) representations of a database entity such as a product or customer. They are also good for storing "sparse" data in general, that is to say irregular (semi-structured) data that would require an extensive use of "nulls" in an RDBMS (nulls being placeholders for missing or nonexistent values).

Document Database Examples

- CouchDB (JSON)
- MongoDB (BSON)
- MarkLogic (XML database)
- Berkeley DB XML (XML database)

Wide-Column (or Column-Family) Stores

Like document databases, Wide-Column (or Column-Family) stores (hereafter WC/CF) employ a distributed, column-oriented data structure that accommodates multiple attributes per key.

While some WC/CF stores have a Key-Value DNA (e.g., the Dynamo-inspired Cassandra), most are patterned after Google's Bigtable, the petabyte-scale internal distributed data storage system Google developed for its search index and other collections like Google Earth and Google Finance.

COLUMN-ORIENTED ADVANTAGE

In row-oriented RDBMS tables, each attribute is stored in a separate column, and each row - and every column in that row - must be read sequentially to retrieve information – a slower method than in the column-oriented NoSQL model, wherein large amounts of information can be extracted from a single wide column in a single "read" action.

These generally replicate not just Google's Bigtable data storage structure, but Google's distributed file system (GFS) and MapReduce parallel processing framework as well, as is the case with Hadoop, which comprises the Hadoop File System (HDFS, based on GFS) + Hbase (a Bigtable-style storage system) + MapReduce.

Primary Uses

This type of DMS is great for:

- Distributed data storage, especially versioned data because of WC/CF time-stamping functions.
- Large-scale, batch-oriented data processing: sorting, parsing, conversion (e.g., conversions between hexadecimal, binary and decimal code values), algorithmic crunching, etc.
- Exploratory and predictive analytics performed by expert statisticians and programmers.

If you are using a MapReduce framework, keep in mind that MapReduce is a batch processing method, which is why Google reduced the role of MapReduce in order to move closer to streaming/real-time index updates in Caffeine, its latest search infrastructure.

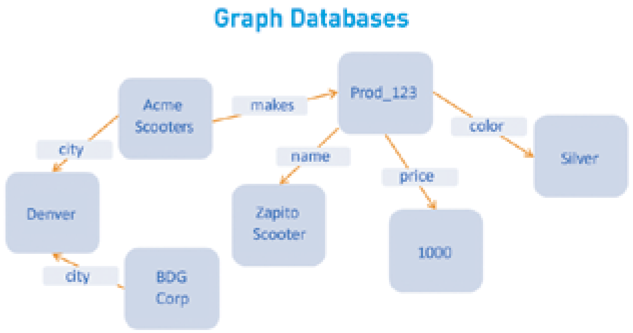
Wide-Column/Column-Family Examples

- Bigtable (Google)
- Hypertable
- Cassandra (Facebook; used by Digg, Twitter)
- SimpleDB (Amazon)
- Hadoop (specifically HBase database on HDFS file system; Apache, open sourced by Yahoo!)
- Cloudera, IBM InfoSphere BigInsights, etc. (i.e., vendors offering commercial and non-commercial Hadoop distributions, with varying degrees of vendor lock-in)

Graph Databases

Graph databases replace relational tables with structured relational graphs of interconnected key-value pairings. They are similar to object-oriented databases as the graphs are represented as an object-oriented network of nodes (conceptual objects), node relationships ("edges") and properties (object attributes expressed as key-value pairs). They are the only of the four NoSQL types discussed here that concern themselves with relations, and their focus on visual representation of information makes them more human-friendly than other NoSQL DMS.

Primary uses



Graph databases are more concerned with the relationships between data entities than with the entities themselves.

In general, graph databases are useful when you are more interested in relationships between data than in the data itself: for example, in representing and traversing social networks, generating recommendations (e.g., upsell or cross-sell suggestions), or conducting forensic investigations (e.g., pattern-detection).

Note these DMS are optimized for relationship "traversing," not for querying. If you want to explore relationships as well as querying and analyzing the values embedded within them (and/or to be able to use natural language queries to analyze relationships), then a search-based DMS is a better choice.

Graph Database Examples

- Neo4j
- InfoGrid
- Sones GraphDB
- AllegroGraph
- InfiniteGraph

Caveats

NoSQL systems offer affordable and highly scalable solutions for meeting particular large-volume data storage, processing and analysis needs. However, the following common constraints should be kept in mind in evaluating NoSQL solutions:

•Inconsistent maturity levels

Many are open source solutions with the normal level of volatility inherent in that development methodology, and they vary widely in the degree of support, standardization and packaging offered. Therefore, what one saves in licensing can sometimes be eaten up in professional services.

• A lack of expertise

There is a limited talent pool of engineers who can deploy and manage these systems. There are likewise relatively few developers or end users who are well-versed in the query languages and tools they use: MapReduce functions (in Erlang, Java, JavaScript, etc.), HQL, Lua, JRuby, SparQL, XQuery, LINQ, JSON/BSON, etc. When available, it can be helpful to choose a commercial, enterprise version of a system – complete with management tools and/or a SQL bridge – to minimize recruiting or outsourcing requirements.

• Inaccessibility

NoSQL systems generally do not provide native full-text indexing (or, consequently, full-text searching), and most do not provide automatic categorization and clustering. A separate search engine would need to be deployed to provide these functions.

• Weak security

In terms of access rights, many have weak to non-existent native security, leaving security to be enforced in the application layer.

In terms of physical security, most compromise on data recoverability in order to boost performance (e.g., Memcached, MongoDB), though most also allow you to manage this tradeoff (e.g., Redis, MongoDB (again), Riak, Cassandra, Voldemort, etc.). Consequently, you should avoid using NoSQL as your primary storage device unless you are certain the system can be configured to meet your particular data durability requirements.

2.NEWSQL

Primary Uses

- High-volume, ACID-compliant OLTP
- Real-time, SQL-based analytics
- Conventional data warehousing & OLAP on Big Data volumes of structured or hybrid data (SQL and, in some instances, MapReduce too)

Definition

Like their NoSQL counterparts, these new (and not-so-new) SQL-based RDBMS achieve Big Data scalability through the use of distributed architectures (in the case of NewSQL, MPP), in-memory processing, the use of solid state drive (SSD) technology and/or by incorporating some NoSQL-inspired flexibility into their data models. Others employ in-database analytics, which is a strategy that combines data warehousing and analytical functions in a single system to reduce latency and avoid the overhead of moving data back and forth between the database and a separate analytics platform.

Those achieving their primary gains through in-memory and SSD technologies tend to be ACID-compliant solutions focused on OLTP. Those gaining a primary advantage through in-database and/or MPP technologies (like inventive parallelization techniques and MapReduce) are generally intended for data analytics and often relax consistency-related ACID constraints to boost performance (exceptions include Oracle Exadata, which supports OLAP and ACID OLTP).

Unlike NoSQL solutions, NewSQL systems tend to be commercial rather than open source (though they may incorporate open source components), with their MPP capacity usually achieved through symmetric processing across a large number of processors embedded within a single high-end computer (usually proprietary), or a small cluster of such computers (VoltDB being an exception).

OLTP-oriented NewSQL systems are ideal for ACID-compliant, high –volume transaction processing in situations where a millisecond can make a critical difference: high frequency trading, battlefield command and control, intrusion detection, network routing, etc.

Depending on the solution, OLAP-oriented NewSQL systems also shine when top speed is critical (e.g., the need to activate real-time triggers or alerts based ton complex analytics), and when you need (or want) to restrict analytics to SQL-based interactions. Some new editions of these platforms also incorporate NoSQL or Search components to enable semi- or unstructured data to be ingested into their analytics infrastructure, with some supporting MapReduce-based analytics as well as SQL analytics. All are (or will soon be) available only in the form of proprietary, integrated software/hardware appliances.

Examples

Examples of systems achieving primary performance advantage through in-memory/SSD technologies (most OLTP-oriented):

- eXtreme DB-64 (embedded db)
- IBM SolidDB
- Oracle TimesTen In-Memory
- Teradata Extreme Performance (OLAP-oriented)
- VoltDB (in-memory)

Examples of systems achieving primary performance advantage through in-database analytics and MPP (most analytics-oriented):

- Greenplum (acquired by EMC)
- IBM DB2
- MS DATAllegro/ SQL Server 2008 R2 Parallel Data Warehouse
- Netezza (acquired by IBM)
- Oracle Exadata (OLTP + OLAP)
- ParAccel Analytic Database
- Teradata Extreme Data
- Vertica (acquired by HP)

Analytics appliances that integrate NoSQL or Search components:

- EMC: Greenplum Chorus (Greenplum Database + Greenplum HD (enterprise Hadoop distribution) + Greenplum Chorus (collaboration layer))
- Oracle: Oracle Big Data Apppliance (Oracle Hadoop distribution + Oracle NoSQL DB (based on Berkeley DB) + other components on Oracle HW, designed to serve as a data source for Oracle 11g or Oracle Exadata (and Oracle Exanalytics, Oracle Exalogic, etc.)
- HP: Idol 10 (Vertica + Autonomy IDOL)
- Teradata: Teradata Aster MapReduce Appliance (Aster MapReduce DB on Teradata HW)

Microsoft is also developing a Big Data appliance that reportedly combines a Microsoft Hadoop distribution with the company's SQL Server and Parallel Data Warehouse software.

Caveats

First, these solutions are expensive. In addition to licensing and development costs, they either need to run on expensive, high-end servers (with the VoltDB exception noted), or they are high-ticket integrated hardware/software applicances, sometimes requiring a rip-and-replace of an existing system, making scaling costly, and restricting business agility through vendor lock-in.

These systems are also expressly engineered for transaction processing or deep analytics. For ACID-compliant transaction processing or complex analytics at Big Data scale, such constraints may represent worthwhile compromises. However, discrete NoSQL and/or Search-based solutions (Cloud or on-premises) are likely a better fit if your needs are more diverse, including, for example:

- Low cost, highly scalable storage of low value data (NoSQL)
- General information search and retrieval (Search or Search + NoSQL for human IR; NoSQL for machine IR)
- Complex exploratory analytics without structured data integration (NoSQL)
- Exploratory analytics for general users (Search)
- Flexible business application development (Search)
- Application-specific data mashups or integrations (Search)
- Low latency – but not sub-millisecond – operational reporting (Search)
- Enrichment of an existing database with unstructured content (Search)

These search-based usages and others are detailed in the next section.

3.SEARCH PLATFORMS

Primary Uses

Processing:

- Natural language processing (NLP)/semantic treatment (text mining; automatic tagging, classification and clustering; relationship mapping, etc.)
- Data aggregation (semantic normalization and integration of heterogeneous data types/sources)

Access/Interaction:

- Full-text, natural language search
- Faceted navigation
- Rapid business application development (customer service, logistics, MRO, etc.)

Analytics:

- Sentiment analysis
- Exploratory analytics (business user)
- Low latency operational reporting/analytics (business user)

Definition

We define a "search platform" as a complete search engine system that can serve as a multi-purpose, information aggregation, access and analysis platform in addition to meeting classic enterprise or Web search needs. Such a search platform, also referred to as a "unified information access" (UIA) platform, encompasses all core data management functions, though with an NLP/indexing twist. These functions include:

- Data capture (crawlers, connectors & APIs)
- Data storage (cached copies of source content and the index itself)
- Data processing (NLP and index construction and maintenance)
- Data access (human and machine IR, faceted navigation and dashboard analytics)

A search system is therefore a DMS like its NoSQL and NewSQL counterparts, and it achieves massive scalability in much the same way, i.e., through distributed architectures, parallel processing, column-oriented data models, etc. However, it is the semantic capabilities and high usability of search-based DMS that make them ideal complements to (and in some cases, alternatives to) NoSQL and NewSQL systems.

First, a search DMS enables full-text search of any NoSQL, NewSQL, or large volume "Old"SQL system (a highly valuable contribution in and of itself). Second, it brings industrial automation to the task of meaningfully structuring data (a must-have for extracting value from Big Data) either for direct use or as a source for another system. A search platform can:

- Effectively structure large volume unstructured content
- Enrich data of any kind with meanings and relationships not reflected in source systems
- Aggregate heterogeneous, multi-source content (unstructured and/or structured) into a meaningful whole

To structure unstructured data, a search platform runs content through NLP processors that consecutively break it down, analyze it, and then enrich it with structural and semantic attributes and values. Take the processing of an HTML page, for example. First, in text-centric processing (see the section on Crawlers in Data Capture & Preprocessing), a crawler captures basic structural information about the page, like page size, file type, and URL, and transmits it along with the page text to an indexer.

The indexer complements this baseline information with the results of semantic analysis to create a holistic "document" to be indexed. At a minimum, this analysis includes a determination of what language the text is written in, followed by parsing the content for indexable keywords (and ultimately phrases), determining along the way the grammatical form of each keyword, and possible grammatical and semantic variants for it. More sophisticated indexers may then analyze the text to identify synonyms and related terms, to flag known people, places or things (using standard or custom lists), to determine the general subject matter treated, to decide whether the overall tone is positive or negative, etc. Business rules may be used to guide the analysis and to perform various types of ETL-style data transformations. This may include extracting only a select number of attributes in order to distill Big Data down into a pertinent and manipulable subset.

Once this structured version of a previously unstructured document has been created, semantic technologies can be used to identify links between it and other documents, whether the other documents are derived from structured sources like databases, semi-structured sources like Web logs, or other unstructured sources like file servers. In this way, you can build a unified, meaningfully organized Big Data collection from any number or type of source systems, and you can further search, explore and analyze this information along any axis of interest (products, people, events, etc.).

When your target user is a business user and not an expert programmer or statistician, the search foundation provides a singular advantage: no other technology is as effective as search at making Big Data meaningful and accessible to ordinary human users.

Tools like natural language search, faceted navigation and data visualization provide users of all skill levels with an instantly familiar way of exploring and analyzing Big Data.
That is to say, they allow a user to launch any search or analytical task the same way they launch a search on the Web: by entering a phrase or a few keywords in a text box. They also enable a user to conduct iterative exploratory analytics simply by clicking on (traversing) dynamic data clusters (represented as text menus or in visual forms like charts or graphs).

This ease of use plus the sheer responsiveness of search platforms encourages iterative exploration: if users get instant answers to questions they ask in their own way, they are enticed to query and explore further. If questions are difficult to formulate and/or answers are sluggish in coming, users will look elsewhere, or give up their quest altogether.

Search platforms are responsive because they are optimized for fast query processing against large volumes of data (read operations), and because most of the calculations they use to produce dashboard analytics and ad hoc drilling are automatically executed as part of routine indexing processes: the results are there waiting to be exploited with no processing overhead (CloudView extends analytic possibilities with high-performance query-time computations).[18]

What's more, all of these out-of-the-box search, access and analysis capabilities can be rapidly packaged into secure, task-oriented business applications to help you extract real bottom-line value out of your Big Data investments in a matter of days or weeks.

For all these reasons, search platforms serve as perfect complements to NoSQL and NewSQL systems, and, in some contexts, provide a pragmatic alternative to them.

Examples

The platforms below are available as standalone systems you can deploy for multi-purpose use in your organization. There are many other platforms which are not sold in standalone form, but rather provide the infrastructure for commercial search-based applications.[19] Others have been absorbed into the integrated HW/SW analytical appliances discussed in the NewSQL section in the wake of acquisitions (see below).

Please note that while all of these platforms are designed for use with large data sets, we can't vouch for their performance with Big Data sets (with the exception of CloudView, of course). We therefore recommend you put their products through the usual screening procedures: references, performance benchmarks for similar deployments, a Proof of Concept (POC) using your own data, etc.

- Attivio Active Intelligence Engine
- EXALEAD CloudView™
- Expert System's Cogito
- Fabasoft Mindbreeze
- Isys Search Software
- Lucene/Nutch/Solr (Apache)
- Sinequa
- Vivisimo Velocity

The search platforms from Autonomy and Endeca have historically been used for SBAs, too (though mainly vertical ones), but in the wake of their recent acquisitions by HP and Oracle, respectively, it appears they will be absorbed into analytic HW/SW appliances:

- The latest Autonomy IDOL release combines Vertica and Autonomy IDOL in one platform for data warehousing and analytics. For the moment, this platform is available on independent hardware, but HP has stated *"the plan over time is to optimize it for HP hardware."* [20]
- Endeca MDEX is being integrated into Oracle Exadata, Endeca Latitude into Oracle BI EE, and Endeca InFront into Oracle ATG Commerce

You'll also notice that familiar names in search like Google, Baidu, Bing, Yahoo! and Ask are absent from this list. This is because most Web search engines are not available as end-to-end search/UIA platforms for commercial enterprise licensing. The exception is EXALEAD, whose enterprise CloudView platform also powers the company's public Web search service.

Google does, however, have an enterprise search offering, the Google Search Appliance, but it is primarily a black box, plug-and-play tool for meeting basic enterprise search needs rather than a complete search DMS. Google is, however, delivering discrete Big Data DMS functions as Cloud services (see the next section, Auxiliary Tools).

Microsoft likewise has an enterprise search offering, MS FAST, but it is designed for use within an MS-specific information ecosystem rather than as a general DMS platform (the same vendor-centric orientation applies to enterprise search tools from IBM, Oracle (as mentioned above) and SAP).

Finally, it should be pointed out that Apache Lucene also has Web roots, though it is not associated with a big name Web search engine. The Lucene search indexer and the Nutch crawler were developed as the two main components of the open source Web search engine Nutch (with the same person, Doug Cutting, having originated Lucene, Nutch and Hadoop).

However, you have to combine Lucene, Nutch, and the Solr search server (or equivalent components) to get a complete search platform. In addition, as these three systems are open source, you'll need some inside expertise – or regular consulting support – to configure, deploy and administer them. This is especially true for search-based applications as these components lack built-in tools for developing and managing SBAs.

Caveats

Search platforms are mature, highly usable solutions for aggregating, accessing and analyzing large volume multi-format, multi-source data. They are also terrific for quickly developing secure, successful business applications built upon such data. They are not the best choice for archival data storage, OLTP or complex or historical OLAP.

It's essential to keep in mind that not all search engines are not created equal. In addition to the screening procedures mentioned above, it is helpful to use a checklist like the one below to ensure a product can support a wide range of information consolidation, access, discovery and analysis needs in Big Data environments.
Does the search platform...

- Collect and process unstructured, structured and semi-structured data?

- Feature an open, standards-based API and connector framework?
- Support true data aggregation in addition to federated search, mashups and metasearch?
- Use semantic technologies to effectively analyze and enrich source data?
- Automatically categorize and cluster content to support faceted search, navigation and reporting?
- Provide a search API or built-in dashboard tools for information visualization and analysis?
- Offer a distributed architecture with parallel processing, or an equivalent architecture, for ensuring satisfactory performance, scalability and cost in Big Data environments?

Ideally, the search platform should also be sufficiently mature to automate essential configuration, deployment and management tasks.

C. Auxiliary Tools

1. CLOUD SERVICES

Primary Uses

There is a Cloud offering available now to meet just about every data management need, including:

- Data acquisition
- Data processing/batch computation
- Data access
- Data analytics
- Data storage

Definition

Inspired by the use of a cloud icon to represent the Internet in computer network diagrams, "Cloud Computing" refers to any information technology service or product delivered via the Internet on a subscription or pay-per-use basis.

Business applications are the most familiar class of Cloud services. Labeled "Software-as-a-Service" (SaaS) solutions, these include well-known offerings like Salesforce and Google Apps. Today, almost every enterprise business software vendor offers a SaaS option for their products. What's important in the Big Data context, however, is that we are likely to see a sharp increase in the number of SaaS offerings that incorporate Big Data sources, like large public databases or social media content. Most of these will be SBAs, for greater context and relevancy.

Another class of cloud solutions is the Infrastructure-as-a-Service (IaaS) category. Companies have also long used IaaS solutions as well, such as remote (often virtualized) hosting of corporate websites, with IaaS offerings now reaching into every corner of IT.

In terms of Big Data, the three most important IaaS offerings are:
1. Data storage,
2. Data processing (computational services), and
3. Data acquisition (also called "Data-as-a-Service," or DaaS).

In the case of data storage, many specialty providers of storage/back-up/recovery solutions as well as general Web services companies like Amazon and Microsoft now offer NoSQL-based solutions to help companies affordably store massive collections of semi-structured or unstructured data for archiving or analysis—data companies might otherwise not even have retained.

> *"Cloud computing and new classes of algorithms will make it possible to keep more transaction detail, keep it longer, and commingle it with other large and very interesting secondary data sets (e.g., phone books and property records)."* [21]

On the processing front, Amazon, Google, Microsoft and others are further enabling companies to use their massive MapReduce-based computing infrastructures to process or analyze these collections.

In terms of data acquisition, commercial firms are offering a growing range of data sets that companies can cross-reference with their own Big Data, or with each other, to yield new views and insights. These largely commercial offerings are complemented by an ever-expanding number of public data sets being published on the Internet by government, educational and scientific organizations.

All these diverse Cloud services are helping organizations of all sizes and types work around the technical and financial barriers to exploiting Big Data.

Examples

Storage services:
- Amazon S3
- EMC Atmos
- Nirvanix
- Google Storage (Labs project)

Computational services:
- Amazon Elastic Compute Cloud (Amazon EC2)
- Google Prediction API & BigQuery (as both were initially offered as part of the discontinued Google Labs program,

Google may choose to commercialize them as domain-specific rather than generic data-crunching services, as with the Earth Builder geo-spatial applications)

Data collections:

- Factual (diverse)
- InfoChimps (diverse)
- Windows Azure Marketplace DataMarket (diverse)
- Hoovers (business)
- Urban Mapping (geographic)
- Xignite (finance)

There are also a number of companies that offer Cloud-based database systems (mainly relational) that are more likely to be used for social or mobile enterprise applications than Big Data storage, processing or analytics. These include:

- Database.com
- Amazon Relational Database Service (RDS)
- Microsoft SQL Azure
- Xeround

Caveats

In addition to addressing concerns common to the Cloud model in general (like privacy, efficiency, vendor lock-in, interactivity, etc.), one needs in particular to carefully weigh the unique challenges of working remotely with very large data sets. Such sets are expensive and slow to move around, and can tax even the best network capabilities.

For example, with a T1 (1.544Mbps) connection, it would take a minimum of 82 days to upload one terabyte of data, and a minimum of two weeks with a 10Mbps connection, which is why Amazon AWS proposes shipping portable storage devices instead, with Amazon then using its high-speed internal network (bypassing the Internet) to get the data to its final Amazon destination. [22]

2. VISUALIZATION TOOLS

Primary Uses

- Reporting & Analytics

Definition

Representing Big Data in visual form helps make it comprehensible to human beings. It is such an effective aid that most science, engineering and Business Intelligence (BI) software features embedded 2D and 3D data visualization and navigational tools.

On the part of major BI vendors – including SAP Business Objects, IBM Cognos, MicroStrategy, SAS Institute, and Information Builders – visualization capabilities include, for example, interactive bar charts, dashboard gauges, pie charts and geographic mapping. SBA engines like CloudView offer this capability as well, generating navigable representations like heat maps, scatter plots, charts, bullet graphs, relationship graphs, tag clouds, sliders, wheels and geospatial maps.

In addition to 2-D and 3-D plotting functions and 3-D volume visualization functions, many visualization tools also include the ability to export results to popular graphics formats.

Examples

Examples of standalone visualization tools include:

- Advizor
- Gephi
- JMP
- Panopticon
- Spotfire
- Tableau

Caveats

Visualization is a terrific tool for summarizing large sets, and for discovering and exploring unexpected relationships and trends, but it's important to select the right type of representation for a given data set and analytical end. Otherwise, you might wind up with a representation that's misleading, confusing or just plain unreadable. To mitigate this risk, and to make the best use of visualization in general, make sure the tool you use produces fully interactive representations.

Keep in mind too that graphical rendering can be a resource-intensive process with very large data sets. In addition, if you want to use a standalone visualization tool, keep in mind that these may need to batch load large data sets into the visualization engine.

5) CASE STUDIES WITH SEARCH

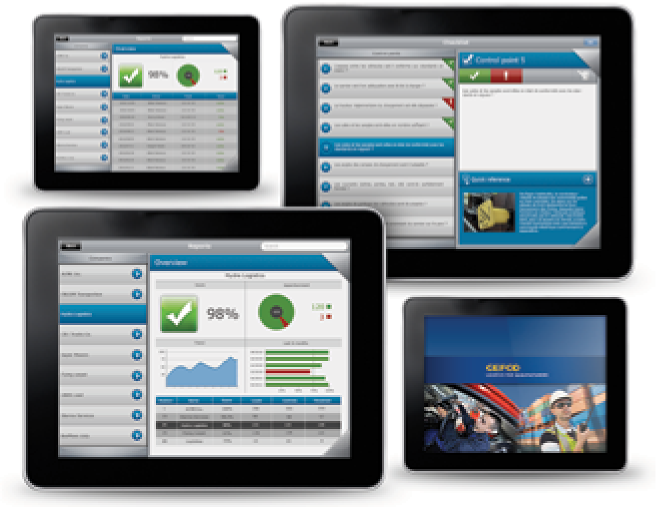
A. GEFCO

Breaking through Performance Barriers

With over 10,000 employees present in 100 countries, GEFCO is one of the top ten logistics groups in Europe. The company provides multimodal transport and end-to-end supply chain services for industrial clients in the automotive, two-wheel vehicle, electronics, retail, and personal care sectors.

The company's automotive division is responsible for the whereabouts of 7 million vehicles on any given day. GEFCO was using an Oracle database to track these vehicles and the 100,000 daily logistical events in which they were involved, and to make this logistical data available to customers through GEFCO's Track & Trace portal.

Several years ago, the Track & Trace portal began to falter under a heavy load. After 2 years of expensive optimization projects, GEFCO was still encountering performance difficulties with the Track & Trace system: complex queries took minutes or even hours to process, data latency was approximately 24 hours, and customer access had to be restricted during business hours to avoid conflicts between information requests and internal transaction processing. At only 3TB, GEFCO had a "Big Data" problem on its hand.



GEFCO: Long, complex forms replaced by a single text box for launching complex queries, with navigation and search refinement supported by cartography and dynamic data facets.

Rather than continuing to ramp up its existing RDBMS, or acquiring a NewSQL system, GEFCO decided to redeploy the Track & Trace portal as a CloudView SBA. The result was an award-winning makeover that boosted performance, enhanced usability, and enabled operational Business Intelligence—at half the cost of the legacy solution.

In addition to slashing per-user costs 50%, the use of an SBA also allowed GEFCO to:

- Cut query response time to a sub-second rate
- Drop data latency from 24 hours to 15 minutes (a rate selected by GEFCO, though the system can support a quasi real-time refresh rate)
- Increase the user base 100-fold—with no end user training (complex forms were replaced by a single search text box and faceted navigation)
- Achieve a 99.98% availability rate with a limited material investment
- Offer customers operational reporting and analytics with visual dashboarding and unrestricted drill-down and roll-up
- Preserve the transactional performance of the Oracle system by offloading IR and analytics

What's more, the initial prototype for this agile application was developed in just 10 days, with the first full production version released iteratively over 3 months.

> *"Every day, I see thousands of events consolidated in real time and I log on to the system just to be sure it's real! The stability and performance of the application is astonishing given its highly innovative character."*
>
> Guillaume Rabier, Director of IT Studies & Projects, GEFCO

The nature of GEFCO's business and the special characteristics of the revamped application (fresh data, instant responsiveness, high availability, and maximum usability) made rolling out a mobile version of the application a natural next step. As the application was already endowed with mobile-ready usability, the main task was to adapt the application for a small screen format and mobile modes of input. Routing and mapping capabilities were then integrated to create a highly successful mobile application for logistics.

B. Yakaz
Innovating with Search + NoSQL

Founded in 2005, Yakaz is a popular vertical Web search engine for classified advertisements (housing, cars, motorbikes, employment and miscellaneous goods and services).The site provides unified access to 60 million ads in 50 languages from tens of thousands of websites, with visitors able to click through to source ads for items of interest. It's a time saving and informative service that draws 15 million unique visitors each month.

The company's founders, two former AOL executives, decided to build their new service on the EXALEAD CloudView^{TM} platform. They were convinced that using CloudView would enable them to go to market quickly and to scale rapidly and massively, and that it would provide an agile foundation that could evolve as their business evolved.

Their faith was well-placed: Yakaz was able to launch the new service in only three months, with deployment handled 100% by their own 3-person staff—and Yakaz achieved profitability in just months.

Yakaz launched first in North America, and the portal was an instant success. Using only 3 commodity servers for its CloudView deployment, Yakaz was soon processing 40 queries per second for a user base that quickly grew from 1/2 million to 6 million monthly visitors. Using those same three servers, Yakaz replicated the service to ten more countries, providing access to more than 10 million ads from 10,000-plus websites in 15 languages.

Today, Yakaz has expanded worldwide, reaching customers in 193 countries, and developed a well-rounded and innovative Big Data infrastructure that now includes four core components:

- EXALEAD CloudView™

CloudView is used to crawl the Web for ads, to automat cally structure the "dirty" Web data the crawler extracts, to process ads submitted via RSS and XML, to build and update the Yakaz index, and to deliver scalable query processing

- Cassandra

An open source Apache project, Cassandra is a NoSQL database that combines Dynamo's distributed design and Bigtable's Column-Family data model. It is used to manage and store user and application data.

- Ejabberd

Ejabberd is a Jabber/XMPP instant messaging server, licensed under GPLv2 (Free and Open Source), and written in Erlang/OTP. It is being used to help Yakaz incorporate social interactions into the portal, beginning with the new user-to-user instant messaging service.

- OpenStreetMap

OpenStreetMap is an open source geo-data service. Its maps are created using data from portable GPS devices, aerial photography, public data sources, and user submissions. It provides the geo-data used by the portal's map-based search and search refinement functions.

This unique infrastructure gives Yakaz the agile, scalable platform it needs to accommodate its rapid growth and innovative spirit. And, though it is not part of the Yakaz business plan, it is an architecture that could support Web-based Big Data analytics. It could be used to reveal the rich information intelligence contained within the tens of millions of ads Yakaz indexes: the average resale value of a particular car make and model, the current state of the rental housing market in a particular location, trends in recruiting – what's hot, what not, and where? It's a long list of possibilities that one could realize with CloudView almost as simply as flipping a switch.

C. La Poste
Building Business Applications on Big Data

France's La Poste Group is Europe's second largest postal operator, with revenues of €20.9 billion in 2010. The group's activities are organized along three main lines of business: Mail, Parcels & Express, and Banking, with a network of 17,000 post offices offering consolidated public access to all of La Poste's products and services.

In 2010, La Poste achieved gains in both revenue and operating profit: an admirable achievement given the general weakness of the economy, and the fact that La Poste faces the same challenges as postal operators worldwide—market deregulation/liberalization and a global decline in letter mail volume.

Part of La Poste's success is due to its innovative use of information technology to boost competitiveness and profitability. This includes a pioneering use of search-based applications (SBAs) to solve long-standing IT challenges and to unleash the potential of its Big Data.

We will look here at two CloudView SBAs La Poste has deployed in its mail division, which represents more than 50% of the group's total revenue:
1. A quasi real-time operational analytics platform
2. A multi-channel customer information system

Operational Reporting & Analytics

La Poste uses a CloudView-powered operational reporting and analytics application to monitor and report on 62 billion events annually involving 180,000 people—in quasi real time. It is, needless to say, a Big Data environment, with:

- 55 million mail pieces treated every day
- 3 to 5 events/treatments per letter
- 300 million records created every day, with a peak of 7000 records a second
- A 21-day data retention requirement
- A 9TB index of 6.3 billion records against 90TB of raw data

The SBA works by aggregating data from diverse sources including business applications, mail-sorting machines (Solystic, Siemens) and video-coding equipment for a global view of mail traffic flow. This end-to-end pipeline visibility enables:

- Timely detection and correction of exceptional events (QoS - Quality-of-Service - analysis)
- Quasi real-time anticipation and optimization of processing and distribution flows
- Multi-dimensional analyses for improved strategic planning

In addition, the ability to aggregate and manipulate this data is also enabling La Poste to develop new premium customer services like a secure, virtual P.O. Box for receiving and storing important documents, SMS push messaging for deliveries, a track-and-trace service for letters, delivery of physical mail via email (mail2email), and, for high-volume commercial clients, complete mail campaign management services.

While La Poste could have used a NoSQL database plus a search engine for this application, the CloudView engine alone satisfied all their data capture, processing, storage, and access needs, and it offered out-of-the-box analytics anyone could use. The CloudView SBA generates dynamic operational reporting on-the-fly against all data facets represented in source systems. Point-and-click simplicity ensures any user can generate reports on his or her own: no training, complex forms, SQL queries – or calls to IT – required.

This usability, and the fact that a quasi-real-time refresh rate was fast enough for La Poste's needs, made the relative cost and complexity of a NewSQL system unnecessary as well.

Multi-Channel Customer Information System (CIS)

This CloudView SBA provided an elegant solution to what would otherwise have been a complex and costly undertaking:

- Providing a single point of access to near real-time data managed in 10 large databases for a 360° view of customers and prospects, and
- Leveraging that unified data layer to support customer interactions across all channels: chat, SMS, Web call back, telephone, mail, email, instant messaging and face-to-face interaction in postal outlets.

By choosing an SBA strategy to meet these needs, La Poste was able to deploy the first operational version of its new CIS in only 90 days—with no impact on existing systems. This initial launch has been followed by an ongoing cycle of iterative releases every 3-4 months, each comprised of 2 or 3 sprints (each given a "Go" or "No go" rating after functional and technical testing).

More than just rolling out new features, this agile development methodology is enabling La Poste to adapt a single solution – its CloudView platform – to meet the needs of five different audiences:

1. Sales information and guidance for Telemarketing Staff
2. Information retrieval, updates, and upsell/cross-sell recommendations for Support Staff (350 operators, 7 call centers)
3. Search and updates for the Back Office (25,000 agents, 3,300 facilities)
4. Operational reporting and analytics for Management
5. Self-service for Customers (businesses and consumers)

In each of these contexts, users benefit from unified data access and Web-style simplicity and speed, with a single text box for launching searches, user aids like fuzzy matching and spelling suggestions, and an average 500-millisecond processing rate. Moreover, this top performance has been achieved at a compelling cost, with a lean footprint and the capacity to scale linearly simply by adding low-cost servers.

D. And Many Others...

Below are a few snapshot profiles of some of CloudView's other Big Data engagements to round out those presented above and in Chapter 2. Your CloudView sales representative can provide you with more information on these and other projects.

Rightmove

Rightmove is the UK's top real estate website, attracting 29 million visitors a month. When Rightmove began to encounter cost and complexity issues with its Oracle database system, they decided to offload information search and access to CloudView.

As a result, in only 3 months, Rightmove was able to:

- Dramatically improve their user experience
- Replace 30 Oracle CPUs with 9 search CPUs
- Slash costs from £0.06 to £0.01 per 100 queries
- Support a peak throughput of 400 queries per second (QPS)
- Achieve a 99.99% availability rate

> *"EXALEAD CloudView^{TM} has allowed the speedy development of advanced search functionality whilst reducing search costs 83%."*
>
> Peter Brooks-Johnson, Rightmove Product Director

France Telecom/Orange

France Telecom/Orange, one of the world's largest telecom companies with more than 217 million customers, chose CloudView for a pivotal role in the modernization of the company's technician support systems. CloudView was used to provide business continuity during the decommissioning of the legacy support system, and to improve efficiency and productivity by giving staff a global view of all relevant service and support information, including:

- CRM data (customer name, address, market segment, customer type, etc.)
- Provisioning information (type of equipment, cable length, line impediments, etc.)
- Network monitoring data (status, performance, loads, etc.)
- Contract data (options, contract period, terms, etc.)
- Technical information (intervention history, technician issues on-site, pending appointments, etc.).

This mission-critical SBA was developed in only 17 consulting days, and is used for tracking performance indicators as well as service delivery.

Consumer Goods Company

CloudView is powering a new communications and collaboration platform for a consumer goods company that provides intelligent, multi-channel access to 175TB of raw data (with 3 million new items added daily) for 35,000 users.

BnF

CloudView is providing the information access infrastructure for the "Gallica" digital library project of the French National Library (Bibliothèque nationale de France, or BnF). To date, BnF has digitized more than one million works, including books, maps, manuscripts, images, periodicals, scores and sound recordings, and made them available to the public via its Gallica Web platform, gallica.bnf.fr.

US Dept of Defense

CloudView is powering a private vertical Web for the U.S. Department of Defense centered on environmental information and issues. It includes data accessible through general public search engines (Google, CloudView, Yahoo!, etc.) as well as deep Web content from government, scientific, industry and commercial databases and applications.

EXALEAD Directory Content Enhancer

The EXALEAD Directory Content Enhancer is a tool that enables online directory publishers to harness the boundless resources of the Web to validate, enrich and extend their own content. First, the platform indexes and analyzes a publisher's database(s).Then, it uses innovative and proprietary techniques to mine billions of Web pages and extract just the right content for that directory. The result is a low-labor, low-TCO method of producing content that is:

- Unique
- Exhaustive (relevant on all business categories)
- Engaging
- Accurate (close to 100%)
- Always up to date

Though producing content of this caliber has always been a competitive differentiator for directory publishers, it has become a business imperative with the release of Google's latest ranking algorithm, Panda (which Google calls a "high quality sites algorithm").This algorithm places a heavy weight on unique, high-quality content.

WHY EXALEAD CLOUDVIEW^{TM}?

EXALEAD CloudView^{TM}'s usability, agility and performance have made it the market leader for search and Search-Based Applications (SBAs). It is also the ideal search platform for Big Data environments, offering:

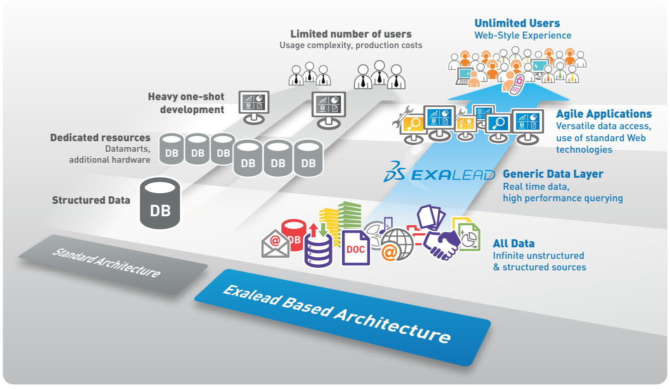
- **Big Data Performance:** Uniquely engineered for both the Web and the enterprise, CloudView can provide advanced semantic treatment of, and secure sub-second query processing against, billions of records for thousands of simultaneous users.

- **Big Data Connectivity:** CloudView features the industry's most advanced Web crawler and state-of-the-art connectors to Big Data sources like message queue data, mainframes, NoSQL stores (e.g., Hadoop HDFS), data warehouses, BI platforms, and social networks.

- **Big Data Analytics:** CloudView's computation and faceting capabilities are the most robust on the market. The platform supports query-time computation of complex numerical, geophysical and virtual aggregates and clusters, and supports dynamic 2D faceting for creating advanced pivot-style tables. Built-in visualization tools, on-the-fly faceted navigation and NLP-based information access ensure anyone can perform exploratory and operational analytics on Big Data, with no training and no calls to IT.

- **Big Data Business Application Development:** Finally, CloudView is unique in providing a drag-and-drop development framework, the Mashup Builder, for rapidly constructing high value business applications on top of your Big Data sources, including applications optimized for mobile delivery.

To learn more about the role CloudView can play in helping you capitalize on Big Data opportunities, we invite you to contact us today for a demonstration of some of our existing large volume SBAs, or to request a Proof-of-Concept (POC) using your own data, a process that usually takes our team just a few days. We guarantee you'll be impressed with the extraordinary value CloudView can reveal in your information assets.


Limited number of users
Usage complexity, production costs
Unlimited Users
Web-Style Experience
Heavy one-shot development
Dedicated resources
Datamarts, additional hardware
DB
Agile Applications
Versatile data access, use of standard Web technologies
EXALEAD
Generic Data Layer
Real time data, high performance querying
Structured Data
DB
DOC
All Data
Infinite unstructured & structured sources
Standard Architecture
Exalead Based Architecture

END NOTES

1. A more literal translation from the Greek is "Give me a place to stand and I will move the Earth," but the often-repeated variation used here more aptly captures the original context — and Big Data challenges and opportunities.

2. For a historical perspective on Big Data-related challenges and technologies, you can survey the proceedings of the annual Very Large Data Base (VLDB) Endowment conferences over the past 35+ years: www.vldb.org.

3. Chris Anderson, Wired Magazine, Issue 16.07, "The Petabyte Age: Because More Isn't Just More — More Is Different," June 2008.

4. 451 Group analyst Matthew Aslett coined the "NewSQL" label in a blog post on April 6, 2011. He applies the label to a recent crop of high-performance, relational SQL databases – mostly open source – including MySQL-based storage engines (ScaleDB, Tokutek), integrated hardware and software appliances (Clustrix, ScalArc, Schooner), and databases using transparent sharding technologies (ScaleBase, CodeFutures). We apply the label to a broader range of evolving SQL-based technologies, and include numerous commercial systems.

5. Gantz J. and Reinsel D., "The Digital Universe Decade – Are You Ready?" IDC, May 2010, Sponsored by EMC Corporation.

6. TheInfoPro Inc. Storage Study, Wave 9, April 2007.

7. See the Wikipedia entry on Lifeloggers: http://en.wikipedia.org/wiki/Lifelog

8. Scott Spangler, IBM Almaden Services Research, "A Smarter Process for Sensing the Information Space," October 2010.

9. McKinsey Global Institute, "Big data: The next frontier for innovation, competition, and productivity," May 2011.

10. Galen Gruman, "Tapping into the power of Big Data," Issue 3, Technology Forecast (Making sense of Big Data), PriceWaterhouseCoopers, 2010.

11. Constance Hays, "What Wal-Mart Knows About Customers' Habits," The New York Times, November 14, 2004.

12. The Economist, "A different game: Information is transforming traditional businesses," February 25, 2010.

13. Jeremy Ginsberg, et al, "Detecting influenza epidemics using search engine query data," Nature, v457, February 2009. See also www.google.org/flutrends/.

14. Alon Halevy, Peter Norvig, and Fernando Pereira (Google), "The Unreasonable Effectiveness of Data," IEEE Intelligent Systems, Issue 2, March/April 2009.

15. Ibid, The Economist.

16. Giuseppe DeCandia, et al, Amazon.com, "Dynamo: Amazon's Highly Available Key-value Store," ACM SOSP'07, October 14–17, 2007.

17. Some content in the NoSQL section is drawn from the book "Search-Based Applications: At the Confluence of Search and Database Technologies," Gregory Grefenstette and Laura Wilber, Morgan & Claypool Publishers, December 2010. For more information on NoSQL systems, see also www.nosql-database.org .

18. Search engines apply counts to every indexable entity and attribute in a source system (references to a particular person in an email system, number of products in a database with a particular attribute, regional sales from an ERP, etc.). These sums feed the statistical calculations used to determine content ranking and relevancy. Search-based applications work by repurposing the results of these counts, calculations and clusters for data presentation, exploration and analytics.

19. Also see "Search-Based Applications," Chapter 10 (citation footnote 17), for additional information on search vendors and search-based products.

20. Development direction outlined by Nicole Egan, chief marketing officer for HP's new information business management group, as reported in "HP yokes Autonomy, Vertica together for Big Data push," GigaOM, Barb Darrow, November 29, 2011.

21. Jeff Jonas, "Sensemaking on Streams," Jeff Jonas Blog, February 14, 2011.

22. From AWS website: "AWS Import/Export - Selecting Your Storage Device," aws.amazon.com/importexport/

Delivering Best-in-Class Products

3DS CATIA
Virtual Product Design

3DS SOLIDWORKS
3D for Professionals

3DS SIMULIA
Realistic Simulation

3DS DELMIA
Virtual Production

3DS ENOVIA
Global Collaborative Lifecycle Management

3DS GEOVIA
Virtual Planet

3DS EXALEAD
Information Intelligence

3DS NETVIBES
Dashboard Everything

3DS 3DSWYM
Social Innovation

3DS 3DVIA
Online 3D Lifelike Experiences

About EXALEAD

Founded in 2000 by search engine pioneers, Dassault Systèmes EXALEAD® provides search and unified information access software that drives innovation and performance in the enterprise and on the Internet. The company's EXALEAD CloudView™ platform is the industry's most sophisticated and scalable infrastructure for Search-Based Applications (SBAs), with over 30,000 business decision makers, half a million enterprise search users, and 110 million Internet users relying on EXALEAD to make their information universe accessible and meaningful.

EXALEAD EMEA
Dassault Systèmes
10 place de la madeleine
75008 Paris
France

Visit us at
3DS.COM/EXALEAD