

Massive Data Analytics and the Cloud

A Revolution in Intelligence Analysis



by

Michael Farber
farber_michael@bah.com

Mike Cameron
cameron_mike@bah.com

Christopher Ellis
ellis_christopher@bah.com

Josh Sullivan, Ph.D.
sullivan_joshua@bah.com

Booz | Allen | Hamilton

strategy and technology consultants

Massive Data Analytics and the Cloud

A Revolution in Intelligence Analysis

Cloud computing offers a very important approach to achieving lasting strategic advantages by rapidly adapting to complex challenges in IT management and data analytics. This paper discusses the business impact and analytic transformation opportunities of cloud computing. Moreover, it highlights the differences among two cloud architectures—Utility Clouds and Data Clouds—with illustrative examples of how Data Clouds are shaping new advances in Intelligence Analysis.

Intelligence Analysis Transformation

Until the advent of the Information Age, the process of Intelligence Analysis was one of obtaining and combining sparse and often denied data to derive intelligence. Analysis needs were well-understood, with consistent enemies and smooth data growth patterns. With the advent of the Information Age, the Internet, and rapidly evolving and multi-layered methods of disseminating information, including wikis, blogs, e-mail, virtual worlds, online games, VoIP telephone, digital photos, instant messages (IM), and tweets, the raw data and reflections of data on nearly everyone and their every activity are becoming increasingly available. This availability and constant growth of such vast amounts of disparate data is turning the Intelligence Analysis problem on its head, transforming it from a process of “stitching together sparse data to derive conclusions” to a process of “extracting conclusions from aggregation and distillation of massive data and data reflections.”¹

Cloud computing provides new capabilities for performing analysis across all data in an organization. It uses new technical approaches to store, search, mine, and distribute massive amounts of data. Cloud computing allows analysts and decision makers to ask ad-hoc analysis questions of massive volumes of data in very quick and precise ways. New cloud computing technologies are driving analytic transformation in the

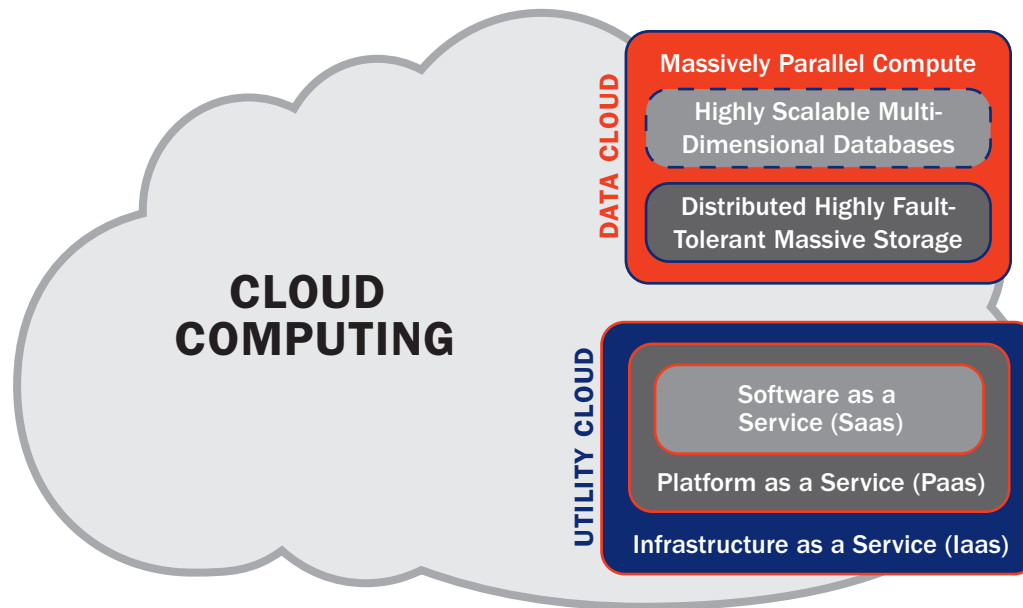
way organizations store, access, and process massive amounts of disparate data via massively parallel and distributed IT systems. These technologies include Hadoop, MapReduce, BigTable, and other emergent Data Cloud technologies. Historically, the amount of processing power that could be applied constrained the ability to analyze intelligence. Consequently, data representation and processing were restricted by the amount of available processing power. Complex data models and normalization became common, and data was forced into models that did not fully represent all aspects of data. Understanding the emergence of “big data,” data reflection, and massively scalable processing can lead to new insights for Intelligence Analysis, as demonstrated by Google®, Yahoo!®, and Amazon®, which leverage cloud computing and Data Clouds to power their businesses. Booz Allen’s experience with cloud computing ideally positions the firm to support current and future clients in understanding and adopting cloud computing solutions.

Understanding the Differences Between Data Clouds and Utility Clouds

Cloud computing offers a wide range of technologies that support the transformation of data analytics. As outlined by the National Institute of Standards and Technology (NIST), cloud computing is a model “for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”² A number of cloud computing deployment models exist, such as public clouds, private clouds, community clouds, and hybrids. Moreover, two major cloud computing design patterns are emerging: Utility Clouds and Data Clouds. These distinctions are not mutually exclusive; rather,

¹ Discussions with US Government Computational Sciences Group.

² US National Institute of Standards. <http://groups.google.com/group/cloudforum/web/nist-working-definition-of-cloud-computing>.

Exhibit 1 | Relationship of Utility Clouds and Data Clouds

Source: Booz Allen Hamilton

these designs can work cooperatively to provide economies of scale, resiliency, security, scalability, and analytics at world scale. As illustrated in Exhibit 1, fundamentally different mission objectives drive Utility Clouds and Data Clouds. Utility Clouds focus on offering infrastructure, platforms, and software as services that many users consume. These basic building blocks of cloud computing are essential to achieving real solutions at scale. Data Clouds can leverage those utility building blocks to provide data analytics, structured data storage, databases, and massively parallel computation, which allow analysts unprecedented access to mission data and shared analysis algorithms.

Specifically, Utility Clouds focus on providing IT capabilities as a service; e.g., Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). This service model scales by allowing multiple constituencies to share IT assets, called multi-tenancy. Key to enabling

multi-tenancy is a security model that separates and protects data and processing. In fact, security—focused on ensuring data segmentation, integrity, and access control—is one of the most critical design drivers of Utility Cloud architectures.

In contrast, the main objective of Data Clouds is the aggregation of massive data. Data Clouds have an architectural approach of dividing processing tasks into smaller units distributed to servers across the cluster with co-location of computing and storage capabilities, allowing highly scaled computation across all of the data in the cloud. The Data Cloud design pattern typifies data-intensive and processing-intensive usage scenarios without regard to concepts used in the Utility Cloud model, such as virtualization and multi-tenancy.

Exhibit 2 contains a comparison of Data Clouds and Utility Clouds. The difference between these two architectures is seen in industry. Amazon is an excellent model for Utility Cloud computing, and Google is an excellent model for Data Cloud computing.

Cloud Computing Can Solve Problems Previously Out of Reach

More than just a new design pattern, cloud computing comprises a range of technologies that enable new forms of analysis that were previously computationally impossible or too difficult to attempt. These business and mission problems are intractable to solve in the context of traditional IT design and delivery models and have often ended in failed or underwhelming results. Cloud computing allows research into these problems, opening new business opportunities and new outreach to clients with large amounts of data.

Much of the current cloud computing discussion focuses on utility computing, economies of scale, and migration of current capabilities to the cloud. Refocusing the conversation to new business opportunities that empower decision makers and analysts to ask previously un-askable questions is the emerging power of the cloud.

What does the cloud allow us to do that we could not do before? Compute-intensive problems, such as large-scale image processing, sensor data correlation, social network analysis, encryption/decryption, data mining, simulations, and pattern recognition, are strong examples of problems that can be solved in the cloud computing domain.

Consider a social network analysis problem from Facebook®. This problem was impossible to solve before the advent of cloud computing:³

With 400 to 500 million users and billions of page views every day, Facebook accumulates massive amounts of data. One of the challenges it has faced since its early days is developing a scalable way of storing and processing all these bytes since historical data is a major driver of business innovation and the user experience on Facebook. This task can only be accomplished by empowering Facebook's engineers and analysts with easy to use tools to mine and manipulate large data sets. At some point, there isn't a bigger server to buy, and the relational database stops scaling.

To drive the business forward, Facebook needed a way to query and correlate roughly 15 terabytes of new social network data each day, in different languages, at different times, often from mixed media (e.g., web, IM, SMS) streaming in from hundreds of different sources. Moreover, Facebook needed a massively parallel processing framework and a way to safely and securely store large volumes of data. Several computationally impossible problems were lingering at Facebook. One of the most interesting of these problems was the Facebook Lexicon, a data analysis program that would first allow a user to select a word and would

Exhibit 2 | Comparison of Data and Utility Clouds

Data Clouds	Utility Clouds
<ul style="list-style-type: none"> • Computing architecture for large-scale data processing and analytics • Designed to operate at trillions of operations/day, petabytes of storage • Designed for performance, scale, and data processing • Characterized by run-time data models and simplified development models 	<ul style="list-style-type: none"> • Computing services for outsourced IT operations • Concurrent, independent, multi-tenant user population • Service offerings such as SaaS, PaaS, and IaaS • Characterized by data segmentation, hosted applications, low cost of ownership, and elasticity

Source: Booz Allen Hamilton

³ Sarma, Joydeep Sen. "Hadoop." Facebook, June 5, 2008. http://www.facebook.com/notes.php?id=9445547199#/note.php?note_id=16121578919.

then scan all available data at Facebook (which grows by 15 terabytes per day), calculate the frequency of the word's occurrence, and graphically display the information over time. This task was not possible in a traditional IT solution because of the large number of users, size of data, and time to process. But the Data Cloud allows Facebook to leverage more than 8,500 Central Processing Unit (CPU) cores and petabytes of disk space to create rich data analytics on a wide range of business characteristics. The Data Cloud allows analysts and technical leaders at Facebook to rapidly write analytics in the software language of their choice. Subsequently, these analytics run over massive amounts of data, condensing down to small, personalized analysis results. These results can then be stored in a traditional relational database, allowing existing reporting and financial tools to remain unchanged but to still benefit from the computational power of the cloud. Facebook is now investigating a data warehousing layer that rides in the cloud and is capable of making decisions and courses of action based on millions of inputs.

The same capabilities inherent in the Facebook Data Cloud are available to other organizations in their own Data Cloud. Cloud computing is a transformative force addressing size, speed, and scale, with a low cost of entry and very high potential benefits. The first step toward exploring the power of cloud computing is to understand the taxonomy differences between Data Clouds and Utility Clouds.

Data Cloud Design Features

The Data Cloud model offers a transformational effect to Intelligence Analysis. Unlike current data warehousing models, the Data Cloud design begins by assuming an organization needs to rapidly store and process massive amounts of chaotic data that is spread across the enterprise, distressed by differing time sequences, and burdened with noise.

Several key attributes of the Data Cloud design pattern are specifically remarkable for Intelligence Analysis and are discussed in the following sections.

Distributed Highly Fault-Tolerant Massive Storage

The foundation of Data Cloud computing is the ability to reliably store and process petabytes of data using non-specialized hardware and networking. In practice, this form of storage has required a new way of thinking about data storage. The Google File System (GFS) and Hadoop Distributed File System (HDFS) are two examples of proven approaches to creating distributed highly fault-tolerant massive storage systems. Several attributes of highly fault-tolerant massive storage systems are key innovations in the Data Cloud design pattern:⁴

- Is reliable, allowing distributed storage and replication of bytes across networks and hardware assumed to fail at anytime
- Allows for massive, world-scale storage that separates metadata from data
- Supports a write-once, sporadic append, read-many usage structure
- Stores very large files, often each greater than 1 terabyte in size
- Allows compute cycles to be easily moved to the data store, instead of moving data to a processor farm.

These attributes are especially important for Intelligence Analysis. First, a large-scale distributed storage system, which leverages readily available commercial hardware, offers a method of replication across many physical sites for data redundancy. Vast amounts of data can be reliably stored across a distributed system without costly storage-network arrays, reducing the risk of storing mission-critical data in one central location. Second, because massive scale is achieved horizontally (by adding hardware) the ability to gauge and predict data growth rates across the enterprise becomes dramatically easier by provisioning hardware at the leading edge of a Data Cloud vice in independent storage silos for different projects.

⁴ Ghemawat, Sanjay; Gobioff, Howard; and Leung, Shun-Tak. "The Google File System." Appeared in 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October 2003. <http://labs.google.com/papers/gfs.html>.

Highly Scalable Multi-Dimensional Databases

The distributed highly fault-tolerant massive storage system lacks the ability to fully represent structured data, providing only the raw data storage required in the Data Cloud. Moving beyond raw data storage to representing structured data requires a highly scalable database system. Traditional relational database systems are the ubiquitous design solution for structured data storage in conventional enterprise applications. Many relational database systems support multi-terabyte tables, relationships, and complex Structured Query Language (SQL) engines, which provide reliable, well-understood data schemas. It is very important to understand that relational database systems are the best solution for many types of problems, especially when data is highly structured and volume is less than 10 terabytes. However, a new class of problem is emerging when dealing with big data of volumes greater than 10 terabytes. Although relational database models are capable of running in a Data Cloud, many current relational database systems fail in the Data Cloud in two important ways. First, many relational database systems cannot scale to support petabytes or greater amounts of data storage, or the database requires highly specialized components to accomplish ever-diminishing increases in scale. Second, and critically important to Intelligence Analysis, is the impedance mismatch that occurs as complex data is normalized into a relational table format.

When data is collected, often the first step is to transform the data, normalize the data, and insert a row into a relational database. Next, users query data based on keywords or pre-loaded search queries and wait for the results to return. Once returned, users sift through results.

Multi-dimensional databases offer a fundamentally different model. The distributed and highly scalable design of multi-dimensional databases means data is stored and searched as billions of rows with billions of columns across hundreds of servers. Even more interesting, these forms of databases do not require schemas or predefined definitions of their structure. When new data arrives in an unfamiliar format, it can be

inserted into the database with little or no modification. The database is designed to hold many unique forms of data, allowing a query-time schema to be generated when data is retrieved, instead of an a priori schema defined when the database is first installed. This is extremely valuable because as new data arrives in a format not seen before, instead of lengthy modifications to the database schema or complicated parsing to force the data into a relational model, the data can simply be stored in the native format. By leveraging a multi-dimensional database system that treats all data as bytes, scales to massive sizes, and allows users to ask questions of the bytes at query-time, organizations have a new choice. Multi-dimensional databases are in stark contrast to today's highly normalized relational data model, which includes schemas, relationships, and pre-determined storage formats stored on one or several clustered database servers.

Moreover, the power of multi-dimensional databases allows computations to be "pre" computed. For instance, several large Internet content providers mine millions of data points every hour to pre-compute the best advertisements to show users the next time they visit one of the provider's sites. Instead of pulling a random advertisement or performing a search when the page is requested, these providers pre-compute the best advertisements to show everyone, at anytime, on any of their sites based on historical data, advertisement click rates, probability of clicking, and revenue models. Similarly, a large US auto insurance firm leverages multi-dimensional databases to pre-compute the insurance quote for every car in the United States each night. If a caller asks for a quote, the firm can instantly tell the caller a price because it was already computed overnight based on many input sources.⁵ The highly scalable nature of such databases allows for massive computational power.

There are some major analytic implications here. First, a computing framework such as a Data Cloud that, by design, can scale to handle petabytes of mission data and perform intensive computation across all the data, all the time means new insights and discoveries are now possible by looking at big data in

⁵ Baker, Stephen. "The Two Flavors of Google." Businessweek.com, December 13, 2007. http://www.businessweek.com/magazine/content/07_52/b4064000281756.htm.

many different ways. Consider the corollaries between Intelligence Analysis and the massive correlation/prediction engines at eHarmony® and Netflix®, which leverage regression algorithms across massive data sets to compute personality, character traits, hidden relationships, and group tendencies. These data-driven analytic systems leverage the attributes of Data Clouds but target specialized behavioral analysis to advance their particular business. Hulu®, a popular online television website, uses the Data Cloud to process many permutations of log files about the shows people are watching.⁵

Massively Parallel Compute (Analytic Algorithms)

Parallel computing is a well-adopted technology seen in processor cores and software thread-based parallelism. However, massively parallel processing—leveraging thousands of networked commodity servers constrained only by bandwidth—is now the emerging context for the Data Cloud.

If distributed file systems, such as GFS and HDFS, and column-oriented databases are employed to store massive volumes of data, there is then a need to analyze and process this data in an intelligent fashion. In the past, writing parallel code required highly trained developers, complex job coordination, and locking services to ensure nodes did not overwrite each other. Often, each parallel system would develop unique solutions for each of these problems. These and other complexities inhibited the broad adoption of massively parallel processing, meaning that building and supporting the required hardware and software was reserved for dedicated systems.

MapReduce, a framework pioneered by Google, has overcome many of these previous barriers and allows for data-intensive computing while abstracting the details of the Data Cloud away from the developer. This ability allows analysts and developers to quickly create many different parallelized analytic algorithms that leverage the capabilities of the Data Cloud. Consequently, the same MapReduce job crafted to run on a single node can as easily run on a group of

1,000 nodes, bringing extensive analytic processing capabilities to users in the enterprise.

Working in tandem with the distributed file system and the multi-dimensional database, the MapReduce framework leverages a master node to divide large jobs into smaller tasks for worker nodes to process. The framework, capable of running on thousands of machines, attempts to maintain a high level of affinity between data and processing, which means the framework intelligently moves the processing close to the data to minimize bandwidth needs. Moving the compute job to the data is easier than moving large amounts of data to a central bank of processors. Moreover, the framework manages extrapolative errors, noticing when a worker in the cloud is taking a long time on one of these tasks or has failed altogether and automatically tasks another node with completing the same task. All these details and abstractions are built into the framework. Developers are able to focus on the analytic value of the jobs they create and no longer worry about the specialized complexities of massively parallel computing. An intelligence analyst is able to write 10 to 20 lines of computer code, and the MapReduce framework will convert it into a massively parallel search—working against petabytes of data across thousands of machines—without requiring the analyst to know or understand any of these technical details.

Tasks such as sorting, data mining, image manipulation, social network analysis, inverted index construction, and machine learning are prime jobs for MapReduce. In another scenario, assume that terabytes of aerial imagery have been collected for intelligence purposes. Even with an algorithm available to detect tanks, planes, or missile silos, the task of finding these weapons could take days if run in a conventional manner. Processing 100 terabytes of imagery on a standard computer takes 11 days. Processing the same amount of data on 1,000 standard computers takes 15 minutes. By incorporating MapReduce, each image or part of an image becomes its own task and can be examined in a parallel manner. Distributing this work to many

⁵ Baker, Stephen. "The Two Flavors of Google." Businessweek.com, December 13, 2007. http://www.businessweek.com/magazine/content/07_52/b4064000281756.htm.

computers drastically cuts the time for this job and other large tasks, scales the performance linearly by adding commodity hardware, and ensures reliability through data and task replication.

Programmatic Models for Scaling in the Data Cloud

Building applications and the architectures that run in the Data Cloud requires new thinking about scale, elasticity, and resilience. Cloud application architectures follow two key tenets: (1) only use computing resources when needed (elasticity) and (2) survive drastically changing data volumes (scalability). Much of this work is accomplished by designing cloud applications to dynamically scale by dynamically processing asynchronously queued events. As a result, any number of jobs can be submitted to the Data Cloud, and those jobs are persistently queued for resilience until completed. The jobs are then removed from a queue and distributed across any number of worker nodes, drawing on resources in the cloud on demand. When the work is complete, these jobs are closed and the resources returned to the cloud.

These features, working in concert, achieve scalability across computers and data volume, elasticity of resource utilization, and resilience for assured operational readiness. These forms of application architecture address the difficulties of massive data processing. The cloud abstracts the complexities of resource provisioning, error handling, parallelization, and scalability, allowing developers to trade sophistication for scale.

The following actions leverage the power of the Data Cloud:⁶

- **Batch Processing Systems**—Log file analysis, nightly automated relationship matching, regression testing, and financial analysis
- **Unpredictable Content**—Web portals or intelligence dissemination systems that vary widely in usage based on time of day, temporal content stores for conferences, or National Special Security Events.

Clearly, applications that can leverage the Data Cloud can take advantage of the ubiquitous infrastructure that already exists within the cloud, elastically drawing on resources as needed based on scale. Moreover, the efficient application of computing resources across the cloud and the rapid ability to decrease processing time through massive parallelization is a transformative force for many stages of Intelligence Analysis.

Conclusion

Cloud computing has the potential to transform how organizations use computing power to create a collaborative foundation of shared analytics, mission-centric operations, and IT management. Challenges to the implementation of cloud computing remain, but the new analytic capabilities of big data, ad-hoc analysis, and massively scalable analytics—combined with the security and financial advantages of switching to a cloud computing environment—are driving research across the cloud ecosystem. Cloud computing technology offers a very important approach to achieving lasting strategic advantage by rapidly adapting to complex challenges in IT management and data analytics.

- **Processing Pipelines**—Document or image processing, video transcoding, indexing data, and data mining

⁶ Varia, Jinesh. Cloud Architectures. Amazon, June 16, 2008. <http://jineshvaria.s3.amazonaws.com/public/cloudarchitectures-varia.pdf>.

About Booz Allen

Booz Allen Hamilton has been at the forefront of strategy and technology consulting for nearly a century. Today, Booz Allen is a leading provider of management and technology consulting services to the US government in defense, intelligence, and civil markets, and to major corporations, institutions, and not-for-profit organizations. In the commercial sector, the firm focuses on leveraging its existing expertise for clients in the financial services, healthcare, and energy markets, and to international clients in the Middle East. Booz Allen offers clients deep functional knowledge spanning strategy and organization, engineering and operations, technology, and analytics—which it combines with specialized expertise in clients’ mission and domain areas to help solve their toughest problems.

The firm’s management consulting heritage is the basis for its unique collaborative culture and operating model, enabling Booz Allen to anticipate needs and opportunities,

rapidly deploy talent and resources, and deliver enduring results. By combining a consultant’s problem-solving orientation with deep technical knowledge and strong execution, Booz Allen helps clients achieve success in their most critical missions—as evidenced by the firm’s many client relationships that span decades. Booz Allen helps shape thinking and prepare for future developments in areas of national importance, including cybersecurity, homeland security, healthcare, and information technology.

Booz Allen is headquartered in McLean, Virginia, employs more than 25,000 people, and had revenue of \$5.59 billion for the 12 months ended March 31, 2011. *Fortune* has named Booz Allen one of its “100 Best Companies to Work For” for seven consecutive years. *Working Mother* has ranked the firm among its “100 Best Companies for Working Mothers” annually since 1999. More information is available at www.boozallen.com. (NYSE: BAH)

To learn more about the firm and to download digital versions of this article and other Booz Allen Hamilton publications, visit www.boozallen.com.

Contact Information:

Michael Farber
Senior Vice President
farber_michael@bah.com
240-314-5671

Mike Cameron
Principal
cameron_mike@bah.com
301-543-4432

Christopher Ellis
Principal
ellis_christopher@bah.com
301-419-5147

Josh Sullivan, Ph.D.
Principal
sullivan_joshua@bah.com
301-543-4611



Principal Offices

Huntsville, Alabama

Sierra Vista, Arizona

Los Angeles, California

San Diego, California

San Francisco, California

Colorado Springs, Colorado

Denver, Colorado

District of Columbia

Orlando, Florida

Pensacola, Florida

Sarasota, Florida

Tampa, Florida

Atlanta, Georgia

Honolulu, Hawaii

O'Fallon, Illinois

Indianapolis, Indiana

Leavenworth, Kansas

Aberdeen, Maryland

Annapolis Junction, Maryland

Hanover, Maryland

Lexington Park, Maryland

Linthicum, Maryland

Rockville, Maryland

Troy, Michigan

Kansas City, Missouri

Omaha, Nebraska

Red Bank, New Jersey

New York, New York

Rome, New York

Dayton, Ohio

Philadelphia, Pennsylvania

Charleston, South Carolina

Houston, Texas

San Antonio, Texas

Abu Dhabi, United Arab Emirates

Alexandria, Virginia

Arlington, Virginia

Chantilly, Virginia

Charlottesville, Virginia

Falls Church, Virginia

Herndon, Virginia

McLean, Virginia

Norfolk, Virginia

Stafford, Virginia

Seattle, Washington

The most complete, recent list of offices and their addresses and telephone numbers can be found on www.boozallen.com