Cloud Computing and Big Data – What's the Big Deal

At one level cloud computing is just Internet enabled time sharing. Instead of organizations investing in all the Information Technology (IT) assets such as hardware, software and infrastructure they need to meet business needs, cloud computing technology makes them available through the Internet. Cloud computing allows an organization to adopt a different economic model for meeting IT needs by reducing capital investments and increasing operational investments. At the same time cloud computing enables the capture, storage, sharing, analysis and visualization of huge amounts of data in multiple formats from multiple media types; Big Data analysis. A recent report from MarketsandMarkets indicates that the global public and private cloud storage market will expand at a compound annual growth rate of 40.2% between 2012 and 2018[1]. The report concludes that the reason for this growth is the emergence of digital trends resulting in the generation of large volumes of unstructured data.

This paper briefly introduces the concepts and implications of cloud computing. Following this the notion of Big Data is introduced and it's relation to cloud computing is explored. Examples will be presented of the uses of big data in industry in general with specific focus on the Aerospace and Defense community.

Cloud Computing

"Cloud computing embraces cyber-infrastructure and builds upon decades of research in virtualization, distributed computing, grid computing and more recently networking, web, and software services".[2] In other words, although the term cloud computing is relatively new, the concepts and technologiesbehind cloud computing have been emerging and evolving for some time. Consumers of cloud computing access hardware, software, and networking capabilities from third party providers in much the same way they get electricity or water from their utility companies.

According to the National Institute of Standards and Technology (NIST) cloud computing delivers five key features [3]:

- On demand self-service – required IT resources are available when they are needed and where they are needed
- Ubiquitous network access – all one needs is an Internet browser and a network connection to get to their applications and data.
- Location independent resource pooling – the location of data centers is irrelevant allowing cloud providers to pick locations where real estate and power are affordable.
- Rapid elasticity – through virtualization and distributed processing the offerings expand and collapse based on user requirements
- Measured service – the utility model requires that there be infrastructure in place to measure the amount of cloud services delivered to consumers.

Chances are good that you are a consumer of cloud computing in some form or fashion. If you have a webmail account such as gmail or Hotmail you are using a browser based system to access the mail applications hosted on the Internet and your messages are being stored in a server in a data center somewhere in the world. The pictures, messages and videos you post to Facebook, YouTube, Twitter, etc. also reside in the cloud.

There are four types of clouds discussed in the literature:

- **Public Cloud** – available to any user of the Internet willing to meet the terms and conditions of the cloud service provider.
- **Private Cloud** – cloud computing infrastructure and technologies are maintained and operated for a specific organization, department or agency. The private cloud is owned by the organization, department or agency that utilizes it. Private clouds are created when an organization wishes to take advantage of cloud computing technology but does not wish to have their data or applications on the Internet.
- **Community Cloud** – Cloud infrastructure that is established and maintained where several organizations, departments or agencies have similar needs and requirements - such as to support an academic collaboration.
- **Hybrid Cloud** - This implies an intermingling of two or more of the types of clouds in a way that they are sensibly integrated into an organization's way of doing business.

In addition to there being several types of clouds, there are also several types of cloud computing offerings:

- Software as a Service (SaaS) – Applications that are accessed via the cloud. End users access commercially available software applications remotely through the Internet. Typical examples include project management, document management, customer relationship management, and social networking applications
- Infrastructure as a Service (IaaS) - Computer infrastructure is accessed via the cloud. Rather than purchasing, provisioning and maintaining servers, data centers, and networking equipment, end users utilize computer infrastructure, generally through a platform virtualization environment, through the Internet. Iaas is usually purchased on a utility computing model basis where the user only pays for the resources that they utilize. Typical examples include backup and recovery, storage, service management, and computation power.
- Platform as a Service (PaaS) – Development platform is accessed via the cloud. The end user has access to the hardware, software and infrastructure necessary to develop or test applications. Typical examples include database, development and testing, and business intelligence environments.

Benefits and Risks

One could easily see the value cloud computing might bring to an organization – particularly small to medium enterprises (SMEs) that may not have the capital to invest in the IT infrastructure that might take their business to the next level. It is quite clear that for many organizations turning to cloud services would result in cost savings. Instead of having to buy and maintain hardware and software they can simply access it through the Internet. They could also reduce their IT footprint resulting in potential cost savings on real estate and power consumption.

Cloud computing certainly offers potential benefits for organizations that experience peaks and valleys in their requirements for IT resources (if you think about it – that pretty much means any organization that doesn't have round the clock operations). According to James Staten of Forrester "Most enterprise data centers are using less than 50 percent of the total capacity of their resources." [4]. Organizations using cloud services need not resource for their peak needs but rather take advantage of on demand scalable services as they are needed.

As cloud services are adopted in an organization, freeing IT personnel from running backups and installing patches, the IT group has more time to devote to aligning IT services with business needs, creating opportunities for more organizational agility. Because cloud providers are in the business of delivering IT services around the clock, they have a great deal of redundancy built into their offerings. Cloud consumers can expect excellent levels of reliability and availability.

Consumers of cloud services have ubiquitous access to their data and applications. They also are offered virtually unlimited storage space and processing power. These facts create an ideal backdrop for the kinds of challenges that Big Data creates.

While it is true there are many potential benefits of cloud computing, not all clouds are the white and fluffy kind; there are also potential risks associated with an organization moving some or all of their IT services into the cloud. Two of the most frequently cited concerns are security and maturity. Security concerns seem obvious; organizations have sensitive data that they don't want to fall into the wrong hands. Concerns over maturity relate to the fact that there aren't consistent standards across the industry so there exists potential portability issues associated with moving from one cloud provider to another. The good news is that both of these concerns are being proactively addressed by the federal government. In 2009 GSA's cloud office established working groups on both security and standards and in 2010 they launched a government wide security certification and accreditation process for solutions in the cloud.

There are also data governance, privacy and legal issues that may be associated with ventures into the cloud. Who owns the data and how do they assure that the data is not made available to nascent third parties who have no business accessing it? If a particular consumers data resides in a foreign country, what set of laws apply if the data is lost or a contract is breached? These are issues that need to be understood and addressed as organizations move important business processes into the cloud. Another challenge is the fact that progress into the cloud leads to loss of control on the ground. IT groups need to relinquish control to an outside source and are often not able to create custom solutions to meet unique business needs.

Big Data and Cloud Computing

Forrester defines big data as "the techniques and technologies that make capturing value from data at extreme scales economical" [5]. Wikipedia defines it as "a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, analysis and visualization" [6] Basically Big Data refers to number crunching of epic proportion, accomplishing in minutes what may have taken weeks several years ago.

So what does this have to do with cloud computing? Certainly the notion of Big Data can exist without cloud computing. The question to ponder is whether the notion of Big Data would have been conceived without cloud computing. The average teenager spends an inordinate amount of time sharing thoughts (text), videos, and photos with their friends via FaceBook, Instagram, Google+, Twitter, Pinterest, etc.In the US in 2011, retail shopping websites earned $162 billion and the number of online shoppers is expected to grow from 137 million in 2010 to 175 million in 2016[7]. The average number of Google searches per day went from 60 million in 2000 to 4.717 billion in 2011[8]. These applications all exist in the cloud and their providers take Orwellian interest in every transaction that is made. How else would Facebook know who we might want to friend and Amazon knows what books to recommend we read?

So cloud computing is certainly an enabling technology for Big Data and has led to vast amounts of data being collected and stored. Add to this the vast amounts of data collected from other sources through applications and other devices designed to collect and transmit data. Now consider the fact that this data is being collected in many formats; text, video, still images, audio, sensor readings, GPS coordinates, radio frequency identification (RIF) readers, etc. are all thrown into the pot. Big Data is the tools and techniques that make it possible to process these large amounts of data in varying formats with lightning speed.

This brings us full circle back to cloud computing. A recent RedHat Report indicates that many businesses implemented cloud based environments last year as a way to manage the influx of structured and unstructured data. [1]. The cloud not only provides storage solutions for the vast amounts of data being collected but also provides enough computing power to make analysis and visualization of the data possible.

Big Data in Practice

Big Data is much more than just a newer faster way to crunch numbers. There is a great deal of speculation that Big Data can be a game-changer in many facets of life. TechAmerica Foundation conducted a study sponsored by SAP of 200 IT Decision makers in Federal and State Public Sector Organizations. Among their key findings were the facts that Big Data could save governments 10% annually and Big Data could be used to save a significant number of lives [9]. The potential applications of Big Data are endless. Some potential uses include threat detection, battlefield analysis,business intelligence, machine reading, medical research, visual intelligence, health care monitoring, weather prediction, etc. Basically any decision process that could be informed by huge amounts of data collected in heterogeneous formats will improve with Big Data tools and techniques.

On March 29th, 2012 the White House announced President Obama's "Big Data Research and Development Initiative" with federal agencies committing $200 million to find ways to significantly improve the tools and techniques needed to analyze vast quantities of heterogeneous data.[10]. The Big Data Fact Sheet [11] highlights many of the federal programs and initiatives seeking to incorporate and improve Big Data technologies.

The National Science Foundation along with 9 other research funding organizations from Canada, the Netherlands, and the UK announced Round 3 of the Digging Into Data Challenge on Feb5, 2013. Digging into Data challenges the research community to help create new research infrastructure to address Big Data challenges. Researchers form international teams submit their proposed research topics and winning teams get two year grants to execute their research. Below is a sample of some of the projects that round 2 awardees are working on:

- The Chartex Project – this team is developing new ways of exploring full text content of digital historical records. Using medieval charters documenting life from the 12th to the 16th centuries the tool will use a combination of Natural Language Processing and Data Mining to extract information about people, places and events automatically and find new relationships between these entities. The project will then create a virtual workbench that will allow historians and archivists to explore the information in a free form fashion far superior to the meta data focused searches currently available[12]

- The Electronic Locator of Vertical Interval Successions (ELVIS): The first large data driven research project on Musical Style – this team is studying the changes in Western Musical style from 1300 to 1900 using digitized collections of several large music repositories. The aim of the project is to study changes in European polyphony over the period using advanced musical information retrieval techniques to compare highly contrasting kinds of music that are unified by common concepts of tonality, consonance, dissonance and voice leading.[13]

These projects are academic in nature and are focused, by design, on social sciences and humanities. They are however an interesting window into the potential of Big Data and the ways that it can advance research across many fields of study. In addition to funding academic research, much federal money is focused on delivering Big Data capability to support ongoing federal initiatives. The Department of Defense has several programs currently underway which plan to use Big Data. Several of these programs are outlined below:

- The Insight Program seeks to develop an adaptable, integrated human-machine Exploitation and Resource Management System (ER&M) which they expect to be the next generation of Intelligence, Surveillance and Reconnaissance (ISR) technology. The Insight program will be providing analysis and exploitation of data from multiple sources including imaging sensors, non-imaging sensors and other sources, directly to the tactical users on the battlefield. Through the application of behavioral discovery and prediction algorithms, battlefield threats will be detected.[14]

- The Minds Eye program seeks to safely provide surveillance capability using visual intelligence. Army scouts are often tasked to enter uncontrolled areas and set up observation posts for temporary surveillance activities. Replacing these soldiers with a truly smart camera would remove them from harm's way. The device being developed could be instructed to report on activities of interest eliminating those activities that are normal or non-threatening. Today's machine vision technology is focused on recognizing a wide range of objects and their properties – the nouns in the scene. The proposed technology differs from this in that it adds the perceptual and cognitive underpinnings to recognize and react to the verbs in those scenes. [15]

These are merely two examples of the strides that the Department of Defense is taking in the area of Big Data. Visit the Big Data Fact Sheet [11] for more examples of how this technology can be exploited to garner improvements for many targets of federal spending.

Conclusion

Cloud computing is an enabling technology for Big Data from several respects. Operations in the cloudopen up vast amounts of opportunities for collection of all types of data by providers of cloud services. Couple this with the fact that cloud services make it possible for researchers to access on demand to nearly unlimited storage capacities and processing power. Cloud computing provides a stage for making Big Data analysis possible.

While there certainly is number crunching involved, Big Data is not merely a new term to describe that number crunching. Big Data refers to the tools and techniques necessary to crunch massive amounts of heterogeneous data in an effort to draw useful conclusions from that data in a reasonable time frame. When the analysis involves identifying threats or detecting network violations, there aren't weeks to wait for the results of analysis.

This final example of a use of Big Data analysis really brings home its power. A researcher at MIT studying natural language development for use in robotics decided to film the first three years of his son's life. He installed audio and video recording devices throughout his home in order to capture all the nuances that lead to his first word, first sentence, etc. From this he ended up with 90000 hours of video and 140000 hours of audio recordings. Using Big Data analysis he and his team were able to turn these hours of recordings into useful lessons about language development.

There is of course a sort of creepy aspect to the whole notion when Big Data and Big Brother seem to be too closely connected. Technology advances have created an environment where large scale data collection and analysis is possible and it is not surprising that enterprising organizations may want to use this data to their advantage. But Big Data is here to stay. More and more organizations and agencies are finding ways to learn more and do more with the ever increasing amounts of data sent their way. This technology enables these organizations and agencies to save money, improve services and save lives.

References

[1] "Big data's impact will influence cloud computing in 2013", Feb 4, 2013 , Digital Reasoning, available at http://www.digitalreasoning.com/2013/industry-news/big-datas-impact-will-influence-cloud-computing-in-2013/ retrieved Feb 28, 2013

[2] Vouk, M.A., "Cloud computing – Issues, Research and Implementations", 30th International Conference on Information Technology Interfaces (ITI 2008), Croatia, 2008

[3] Wyld, D., "Moving to the Cloud: An Introduction to Cloud Computing in Government, IBM Center for the Business of Government E-Government Series, 2009

[4] Leavitt Communications, "Is Cloud Computing Really Ready for Prime Time?", available at http://www.leavcom.com/ieee_jan09.htm (retrieved 4/2011)

[5] McCarthy, Bede, "Big Hopes for Big Data", The Future of Business Data, Feb 18, 2013, available at http://www.ft.com/intl/cms/s/2/1a780a6c-648b-11e2-934b-00144feab49a.html(retrieved 2/28/13)

[6] Big Data, http://en.wikipedia.org/wiki/Big_Data

[7]Kril, Jessica, "Statistics and Facts about Online Shopping", Statista (partner of DowJones) available at http://www.statista.com/topics/871/online-shopping/ (retrieved 3/1/2013)

[8]"Google Annual Search Statistics", Statistic Brain, available at http://www.statisticbrain.com/google-searches/(Retrieved 3/1/2013)

[9]"Big Data and the Public Sector", TechAmerica Foundation, available at http://www.techamericafoundation.org/content/wp-content/uploads/2013/02/SAP-Public-Sector-Big-Data-Report_FINAL.pdf(retrieved 3/4/13)

[10].Kalil, Tom, "Big Data is a Big Deal", Office of Science and Technology, March 29,2012, available athttp://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal(retrieved on 3/1/2013)

[11]"Big Data Across the Federal Government", Executive Office of the President, March 29, 2012, available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf(retrieved 2/28/2013)

[12]Chartex project website, available at http://www.chartex.org/(retrieved 3/1/2013

[13] Digging into Data Website, ELVIS page available at http://www.diggingintodata.org/Home/AwardRecipientsRound22011/ELVIS/tabid/192/Default.aspx(retrieved 3/1/2013)

[14]DARPA Website, Insight webpage, available at http://www.darpa.mil/Our_Work/I2O/Programs/Insight.aspx(retrieved 2/26/13)

[15] DARPA Website, Mind's Eye webpage, available at http://www.darpa.mil/Our_Work/I2O/Programs/Minds_Eye.aspx(retrieved 2/26/13)

