Big Data Status
By Jaqui Lynch

Introduction

Big Data really burst onto everyone's horizon around 2012, when the sheer amount of data that organizations were trying to manage became unwieldy. The move to look at and analyze new types of data is critical to the ability to maintain competitiveness in today's business world. Understanding the types of data and how best to use and analyze that data can make a huge difference in how a business moves forward.

In 2012 we started to see the initial attempts to integrate significant amounts of unstructured data with the standard business level structured data. The unstructured data came from multiple sources that had not previously been mined and these include social media, web data, intelligent devices and sensors. As an example, the increase in the use of smart fitness devices has led to a plethora of data being uploaded to those vendor's sites. This meant that the initial focus was on volume and variety. Volume refers to the total amount of data that needs to be stored and analyzed and variety refers to the multiple types of data that are being recorded. This data can be traditional structured data or aggregations of unstructured data coming from multiple sources.

Initially, with volume and variety being the focus, we all concentrated on designing a scalable infrastructure that could grow as needed. This meant taking advantage of virtualization on servers and storage and looking at high capacity warehouses that could handle both growth and performance as well as the long term storage of data with rapid recall. But, even more important than that was figuring out how to integrate all this data so that it could be analyzed and viewed in a meaningful manner.

Once these scalability and variety issues were resolved the focus moved to velocity and veracity which, to some extent, is what we are dealing with today. Velocity refers to the speed at which data accumulates or flows into the data center – this could be streaming data or it could be received in multiple other methods, even tape. Velocity requires an agile and flexible infrastructure that has sufficient power to allow real time analysis of enormous amounts of data.

Veracity is more difficult, as it refers to how trustworthy the data being received is. Veracity for streamed data is a key challenge as it is not always possible to determine the exact source or validity of the data received. If you have spent time on social media or on the web you understand that the data needs to be verified. If business decisions or life altering decisions are going to be made based on data, then this raises the issue of trustworthiness of the data to a very high level.

The above four items – volume, variety, velocity and veracity – are often referred to as the four V's of big data. They impact the storage needed as well as the

CPU and network bandwidth required. However, many argue that there is a fifth V – value. Having petabytes of data comes at a significant cost. There is the cost of gathering or receiving the data, storing it, analyzing it and then determining what to do with it in the long term. But if it is just a bunch of bits and bytes with no meaning for the business then there is no point in having that data. Raw data is just that – data with no context and not all of that data has value, although it is possible that the usefulness of some data may change over time. Strategies need to be put in place to determine how to categorize data so that its value can be assigned. The key is collecting all the data and providing it with enough context to be valuable. This is the role of analytics software; adding the context needed to assign value to the data.

Where are we now?

As we saw the focus for big data has shifted from volume to velocity. Software and strategies have been available for some time to deal with the sheer amount of data. Our focus now is on how we take that raw data and give it value using searches, queries, analysis, visualization and other techniques to turn it into useful business information. This also involves mining for relationships between different types of data that may come from multiple disparate sources. This needs to be done quickly and on great volumes of data. To assist with this the world of analytics now includes RAM-based noSQL technologies. These allow developers to create simple data stores for fast access table lookups.

One such example is the IBM Data Engine for NoSQL. This delivers high speed access to both RAM and Flash storage which can reduce costs and improve performance for noSQL platforms. Taking advantage of the POWER8 CAPI (coherent accelerator processor interface), flashmemory can be attached and deliver up to 40TB of extended memory with now performance degradation. This allows for the rapid analysis of enormous amounts of data. Per the IBM web site: “The CAPI technology in POWER8 processor-based servers introduces a new tier of memory for NoSQL data applications. IBM Data Engine for NoSQL can be delivered as a single system, and scale out as needed to support data growth. This translates to lower hardware, maintenance and energy costs without sacrificing performance.” Once the platform is in place then there are multiple software offerings to deal with the data.

New Roles

The move to big data has added a lot of extra work onto the CIO. Someone has to look after all this new data and they also have to decide what to do with it, how to store it and how to protect it. They also need to ensure the data os processed correctly so that it is made useful to the business. This has led to the creation of a new role – the data scientist. The data scientist is a hybrid of science and business as they need to have strong skills in statistics, modelling and mathematics, but they also need to have a good understanding of the business needs and the ability to communicate the processed data to the business in a

meaningful way. More companies are looking for people with these skillsets and they are not always easy to find.

Why bother with Big Data

Big data is about taking data and turning it into information that can be used to make the business more efficient and more profitable. Organizations that use big data technologies broadly throughout their business functions can take advantage of capabilities that enable business functions to consume the data rather than just absorb it. This allows them to are create the greatest impacts on business performance. The organizations that take the time to design an dimplement an agile and flexible infrastructure that is designed to manage data efficiently and move it through the analytics process quickly are far more likely to be able to create business value from their data. This allows them to become more competitive in their chosen markets. The ability to draw relationships from disparate data also allows for far superior planning. This is particularly important in areas like healthcare where the ability to take client medical information and combine it with known weather patterns or local health issues can make a huge difference in a patient's recovery. There are also significant implications for the use of big data in business, law enforcement, fraud detection and many other areas beyond just medical.

Summary

Organizations are now making heavy use of big data technologies throughout their business functions. The ability to process these huge amounts of data is impacted by the network, storage and server scalability. In particular, the sheer amount of CPU bandwidth and memory needed to process the data is a major factor in design choices. Whether you are looking at Hadoop, BigInsights, Infostreams or any of the other multitude of offerings in the big data arena, keep in mind that the underlying memory and I/O infrastructure are going to be critical to your ability to grow the environment and specifically the data that is being ingested, analyzed and stored. Data is no longer the traditional fixed structure that we have used for so long – it now consists of many types of data that need to be carefully integrated and evaluated in order to provide insight for the business. Focus should be on designing solutions that are not only scalable, but that are also easy to use. This allows every decision maker in the company to easily find, analyze and share the information they need access to in order to make the best decisions for the company.

References

IBM Data Engine for NoSQL
http://www-03.ibm.com/systems/power/solutions/bigdata-analytics/data-engine-nosql/

IBM Big Data Resources

http://www-03.ibm.com/systems/power/solutions/bigdata-analytics/resources.html

IBM Big Data Analytics
http://www-03.ibm.com/systems/power/solutions/bigdata-analytics/index.html