Let's Analyze some Data

Data analytics has become a key element of the business decision process over the last decade. Classic reporting on a dataset stored in a database was sufficient until recently, but yesterday's data gathering and mining techniques are no longer a match for the amount of unstructured data and the time demands required to make it useful. The common limitations for such analysis are compute and storage resources required to obtain the results in a timely manner.

The New Era of Big Scale.

Twenty years ago, IT teams were focused on obtaining optimal performance from the key applications and infrastructure of their enterprises. These silos acting as "systems of record" typically did a relatively good job of keeping track of vital information, but they were very expensive, difficult to manage and did not offer sufficient 'drill-down' insight into the data to drive business advantage. Ten years ago, the IT focus shifted to efficiency, or 'how to do more with less'. Technologies like virtualization, sharing, and consolidation of the enterprise's existing infrastructure became the key drivers for IT teams.

Presently, we are entering a new era of big scale, where the amount of data processed and stored by enterprises is breaking down every architectural construct in the storage industry today. As a result, IT teams are trying to convert these existing systems of record, designed and built back in the 1990s and 2000s, into "systems of engagement", which are defined as systems that can efficiently deliver the necessary information, to the right people, in real time, to help them perform more sophisticated analyses and make better business decisions.

Furthermore, this massive increase in scale is occurring for a number of reasons. Due of cost pressures, many companies are consolidating their physical data centers, as they can no longer afford for each business unit to have its own IT infrastructure distributed around the globe. The move to cloud computing also contributes to increased scale, aggregating the demand of hundreds of thousands of users onto fewer, centralized systems.

Another source of the increase in scale is the massive growth in machine-generated and user-generated data. Digital storage technologies are moving to denser media, digital photography is ubiquitous, videos are using higher resolution, and advanced analytics require far more data, and hence more and denser storage. Machine-generated data from sensor networks, buyer behavior tracking, and countless other sources contribute to much larger datasets that must be understood and commercialized. In short, the amount of data is increasing and the data objects themselves are getting bigger. All of these forces together put an enormous amount of scale pressure on existing infrastructures, especially the storage platform.

Big Data Requires Big Plans.

Explosive data growth is a reality and its trajectory is rising quickly. In order to accommodate and support this level of intensification, more robust and powerful data management solutions are becoming increasingly important. Data generation and the diversification of data use drive the adoption of more role-based storage solutions within the data center. These factors, coupled with the transition to highly virtualized data center environments, affects how organizations buy and manage server, storage, and network assets and are key drivers in what is propelling Big Data into an everyday reality. The outlook is Big Data in the Cloud.

Big Data is comprised of datasets that grow so large that they become cumbersome to manipulate using traditional database management tools. Difficulties include capture, storage, search, sharing, analysis, and visualization. The growth trend continues because of the significant benefits of working with larger and larger datasets that allow analysts to discover business trends and solve problems. Though a moving target, current limits are on the order of terabytes, petabytes, and exabytes of data. At this trajectory, even zettabytes (1,000s of exabytes) will be a reality in the not-too-distant future.

Data is everywhere, whether users, applications, or machines create it and it's growing exponentially with no vertical market or industry being spared. Due to this reality, IT organizations everywhere are forced to come to grips with storing, managing and extracting value from every piece of it -– as inexpensively as possible. This begins the real race to cloud computing where the framework needs the ability to process data increasingly in real-time and in far-greater orders of magnitude -– at a fraction of what it would typically cost.

Big Challenges.

Ultimately, today's enterprises find it difficult or impossible to manage the exponential growth in big data. Traditional approaches can't scale to the level needed to be able to ingest all of the data, analyze it at the speed at which it arrives, and store the relevant datasets efficiently for extended periods of time. The industry as a whole has started to get a handle on how to manage the increased infrastructure complexity in a virtual world, but handling infrastructure in a scalable world presents some very serious challenges.

Time-to-information is critical for enterprises to derive maximum value from their data. If it takes weeks or months to run an analysis, it may not be timely enough to detect patterns that may affect the business in an instant. Compliance is also a significant challenge for many enterprises. Regulated organizations may have to keep data for very long periods of time – or forever. In addition, they are required to find the data quickly when needed for reporting, litigation events or during industry audits. Therefore, the challenges of Big Data are all about gaining business advantage, and specifically, how to obtain the most value for the enterprise from this immense digital universe of information. It's also important to be aware of the fact that Big Data is breaking today's storage infrastructure along three major facets:

Complexity. Data is no longer just about text and numbers; it's also about real-time events and shared infrastructure, and the inherent relationships in the data. The information is now linked, is high fidelity, and consists of multiple data types, many of which are unstructured. Applying typical algorithms for search, storage, and categorization is becoming much more complex and inefficient.

Speed. How fast is the data coming in? High-definition video, streaming media over the Internet to player devices, slow-motion video for surveillance, social media streaming feeds – all of these have very high ingestion rates. Businesses have to keep up with the data flow to make the information useful. They also have to keep up with ingestion rates to drive faster business outcomes – or in the military, to save lives.

Volume. All collected data must be stored in a location that is secure and always available. With such high volumes of data, IT teams must make decisions about what is "too much data." For example, they might flush all data each week and start all over the following week. But for many business units and their applications, this is not an option, so more data must be stored longer – without increasing the operational complexity. This can cause the infrastructure to quickly break along the axis of volume.

Home is Where the Data is.

Storage providers play a critical role in the explosive data growth and increase in scale. After all, they store the data and they need to be able to provide a robust enough environment and solution offering to accommodate such datasets. The most effective solutions are ones that efficiently process, analyze, manage, and access data at scale. Specifically, solution portfolios that are organized by the primary use cases of *analytics*, *bandwidth*, and *content* – "ABC" for short - are those that address the key bases for success.

Analytics for extremely large data sets focus on providing efficient analysis for those datasets that are significantly larger than any we've been accustomed to in the past, especially unstructured data. Analytics is all about gaining insight, taking advantage of the digital universe, and turning data into high-quality information, providing deeper insights about the business to enable better decisions.

Bandwidth is related to the performance for data-intensive workloads. High-bandwidth applications include high-performance computing (HPC) and the ability to perform complex analyses at extremely high speeds. They also include high-performance video streaming for surveillance and mission planning as well as video editing and play-out in media and entertainment. Unlike legacy applications where low-latency, low-bandwidth solutions sufficed, data-intensive computing requires high bandwidth.

Finally, *Content* focuses on the need to provide boundless secure scalable data storage. Content solutions must enable storing virtually unlimited amounts of data, so that enterprises can store as much data as they want – but also find it when they need it.

Thinking Bigger & Different.

There are a lot of new and different facets to Big Data. What makes Big Data different is that companies are realizing that all the data they have collected as part of their business operations and all the data that is constantly being collected by video surveillance, web trends, mobile phones, consumer behavior, social media and so on can be combined in interesting and useful ways to gain competitive advantage or have better outcomes. Outcomes spanning a wide range from providing better customer experiences and building better products faster to locating terrorist activity are all centered around Big Data.

Another difference is that most of the data growth that comprises Big Data is unstructured. The simplest example is to compare a customer record that is structured to a video that is unstructured. A customer record has fields like customer name and customer address, it has fixed size, you can store it in a structured (row-column) database, you can search for a specific customer using a query and so on. By contrast, a video is a stream of digital data typically stored as a file. It doesn't have fixed fields and it's difficult to search, therefore it's unstructured. As an example, the opportunity is to be able to store hours, days, months and years of surveillance video, link structured, fixed-field data to it and be able to find the whereabouts and actions of a single person immediately upon request such as identifying a terrorist as soon as they enter an airport or finding known cheats as they enter a casino.

Other big data examples involve gaining insight from very large data sets to identify trends and match them to real-time events. For example, being alerted to a particular customer ordering 300 times more that they usually do so that you can re-route inventory to satisfy their need. This requires analytics to know what they normally order and real-time alerts to events that are abnormal. Similarly banks and brokerages need pattern recognition in real-time to detect fraud.

Many other use cases exist. Consider large retailers analyzing their transactional data together with weather forecasts to anticipate where show shovels need to be delivered ahead of a storm or where fans need to be delivered ahead of a heat wave. There are dozens of other use cases in retailing and merchandising alone.

The drive to develop and deliver innovative products and services in the future will be fueled increasingly by companies' ability to acquire and analyze vast amounts of structured and un-structured data. Large and small enterprises are racing to acquire this capability by leveraging the vast computing power of the public cloud and by re-engineering their data centers into private clouds.

The buzz is real and the challenges are complex, but the fact of the matter is that substantial data growth is everywhere and traditional approaches don't scale (enough). Technology advancements and complexities in model accuracy, real-time information sharing, high-end imaging, streaming video, analytics and other data-intensive applications dramatically are changing the way business is conducted. The time is now to provide robust solutions to manage, support, and maintain these businesses and their *big* data.

Analytics & Big Data is a new committee that has been formed within the Storage Networking Industry Association. The SNIA ABDC is dedicated to fostering the growth and success of the market for what is generally referred as Analytics and Big Data and more generally the use of data storage resources and services by analytics and big data applications and toolsets. The goals of the ABDC are to become the recognized authority regarding the use of storage and storage networking for Analytics and Big Data.

Further goals are to determine and document the characteristics of Analytics and Big Data offerings, the impact of Analytics and Big Data on enterprises and analytics and big data computing as well as collecting requirements from Analytics and Big Data vendors and document best practices in this area. Additionally, the ABDC will collaborate with academia and the research labs of member companies to understand how advances in storage, storage networking, and other technologies will affect Analytics and Big Data.

Furthermore, the ABDC will educate the vendor and user communities on the use of storage and storage networking for Analytics and Big Data. Specific activities proliferating this will be coordinating education activities with the Education Committee, creating peer reviewed vendor-neutral SNIA tutorials and vendor-neutral demonstrations, leveraging Storage Networking World [SNW] and other SNIA and partner conferences, as well as collaborating with industry analysts.

The ABDC will perform market outreach that highlights the virtues of storage and storage networking for Analytics and Big Data such as articles in trade magazines, whitepapers, press releases, and collaborative published articles with academia and research institutions. This committee will collaborate with other industry associations via SNIA's various strategic alliance partners on analytics and big data related technical work in which they are involved. It will coordinate with SNIA Regional Affiliates to ensure that the impact of the Analytics and Big Data Committee is felt worldwide. To promote a well-rounded approach and integration, the ABDC will coordinate with the Cloud Storage Initiative to jointly message the Analytics and Big Data cloud-oriented market and offerings.

Now in its 9th year, and again expected to draw more than 250 developers and engineers, the Storage Developer Conference (SDC) is the only event created by storage developers for storage developers. At this same event, the ***Analytics & Big Data Summit*** *will be held* ***September 20, 2012*** *at the* ***Hyatt Santa Clara****, so don't miss it!*