Datacenter Virtualization:

New Visibility Gaps Create Big Data Opportunities

WHITE PAPER

splunk>

Abstract:

The emergence of new data collection, processing and storage technologies has transformed data conversations from "what do we absolutely need to store" to "what can we do now with more data." Virtualization, arguably the most significant datacenter trend of the last decade, has introduced incredible new efficiencies but brought with them a number of unanticipated technology dependencies and operational challenges. As more and more organizations look to virtualize their mission-critical applications, a clear pathway to addressing these challenges lies in harnessing the data generated by all the different interacting technologies (virtualization, network, storage, etc.). However, such datasets generated by virtualized datacenters define big data in terms of volume, velocity and variety. Managing, monitoring and providing analytics on these environments requires getting the right information at the right time and a big data approach.

This paper discusses the ways in which virtualization and big data are inextricably linked and suggests that only solutions capable of ingesting and exploring large volumes of data in various formats should be considered for monitoring, planning and analytics of virtualized infrastructures.

The Need for Big Data Analytics

Over 11 years ago, Doug Laney, VP of Research Analytics and Performance Management at Gartner, authored a publication[1] on controlling data **V**olume, **V**elocity and **V**ariety. This framework effectively defines big data today. Big data is characterized by **Volume** increasing at incredible rates. There is also a great deal of **Variety** in this data—from web interactions, video, images, user experience, applications, infrastructure response and operations all typically generated by machines. Formats range from completely unstructured, to semi-structured to variously structured and these datatypes oftentimes change. When access to data is critical, the **Velocity** with which data is available for use becomes time-sensitive.

During a recent presentation[2], Jay Parikh, VP of Infrastructure Engineering at Facebook, noted: "Big data is really about having insights and making an impact on your business. If you aren't taking advantage of the data you're collecting, then you just have a pile of data, you don't have big data."

In the last decade, organizations have gone from simply amassing standard data to nonstandard, unstructured forms; from static data to dynamic real-time data. This paradigm shift demands enterprises to alter the way they think. It is not enough to simply capture these large datasets. Solutions focusing on big data need to offer the capability for rich, ad hoc and intelligent analytics on both structured and unstructured data, thus providing new insights and visibility across the entire infrastructure for IT and the business.

Why Virtualization Is a Big Data Problem

Virtualization brings together previously siloed applications, infrastructure and services via software, producing a massive collection of machine data that ranges across users, applications, hypervisors, operating systems, storage, network and physical hardware. All of these layers produce large volumes of data in a variety of formats.

Traditional virtualization toolsets focus on specific vendor technology silos rather than the entire technology stack. As a result, they are unable to provide complete visibility across physical and virtual infrastructures. For example, a storage management tool cannot provide an answer to an application performance problem, because the inter-linkages are not apparent across silos.

With every layer in the virtualized IT stack producing massive volumes of machine data of tremendous variety and velocity, there is a need to manage, easily access and analyze this data for troubleshooting, performance management, capacity planning, security and more. The opportunity is significant and sounds very much like the definition of big data referenced above.

Three Key Capabilities for Leveraging Big Data

In a recent blog post[3], Forrester Research Analyst Mike Gualtieri indicated three important factors that are required to handle big data:

- The ability to store large volumes of data persistently or transiently
- A means to enrich, calculate and analyze this data
- The power to search, query and visualize the data

Let us look at each of these in a little bit more detail.

1. **Storing large volumes of data:** The traditional RDB was not designed to manage or store such large volumes of data streamed in real time, let alone historical data needed for troubleshooting. The machine data generated by all of the systems and infrastructure running in a datacenter is massive. Relational database systems (RDBs) are about structure – storing data in a predictable manner using rows and columns. Delivering this structure requires normalization – the filtering or parsing of data to meet the strict requirements of a RDB. Defining schemas and filtering data with normalization limits the huge opportunity machine data represents. Further, this data is unstructured or semi-structured, has various formats and is real time.

The traditional RDB was not designed to manage or store such large volumes of streaming, let alone historical, data needed for troubleshooting, long-term trending and pattern analysis. Effective management and monitoring can be inhibited by a backend database with pre-defined schemas and constraining data models. Machine data needs to be stored in full fidelity and in raw format, for utilization/analysis by a flexible approach.

2. **Analyzing structured and unstructured data:** When data is stored in its raw format, analysis on such large volumes becomes very resource-intensive, primarily on the CPU and memory. Therefore ad hoc reporting becomes challenging. While an obvious solution to this problem may be to consolidate and aggregate the data and roll it up into averages, thus providing quick results, this compromises data fidelity. For instance, when troubleshooting performance degradations or system availability, aggregation abstracts away important data and inhibits troubleshooting efficiency. One needs the ability to aggregate data for speed, but retain the details for ad hoc analyses.

   Analyzing aggregated non-real time data does not provide the much needed visibility for business-critical applications. Combining real-time metrics with non-aggregated historical data gives valuable week-over-week/day-over-day trends and patterns and not just a point-in-time metric.

3. **The power to search, query and visualize the data:** As we have noted, relational databases are not designed for the complexity, scale or rigidity of today's datacenter. Capturing and storing data is not sufficient if it cannot be rapidly explored, understood and rationalized. Beyond the scalability limits of traditional approaches, there are tremendous flexibility limitations imposed by the RDB approach that prohibit ad hoc analysis and data exploration.

   To fully deliver on the wealth of analytic information hidden within their machine-generated data, organizations need access to interactive charts, tables and graphs that are capable of drilling down into deeper levels of the data set(s). The ideal solution should support any kind of "needle in the data haystack" query, especially when there is not an obvious answer. Such solutions need to provide a rapid means to easily navigate, search, query and visualize the raw data on-demand. Traditional approaches, with rigid schemas and prebuilt queries for "known questions," don't support this important capability and opportunity for data analysis.

Even though virtualization brings about exciting new dynamics to infrastructure management, it creates new challenges in effectively managing and planning the physical and virtual infrastructure. Gaining insight into your virtual deployment and making essential correlations with the applications and other parts of the infrastructure is vital to efficiently managing your resources and gaining the benefits of virtualization.

When Trends Collide – Big Data Meets Virtualization

When we take these big data concepts and combine them with virtualization, the answers to virtualization challenges become more obvious. Virtualized datacenters produce massive volumes of data and there is an inevitable need to **capture and store** this data in full fidelity, **search and analyze** this data easily, and **trend and report** historical and real-time metrics together. Additionally, there is an innate need to **relate data from the various tiers** of virtualization such as hypervisor, guest and host machines, operating systems (or even other virtual IT tiers like network, storage, applications, desktops) with the physical tiers such as storage, network, applications and various other distributed data sources to gain a holistic perspective.

Solutions built for virtualization in the context of big data must provide:

- An easy means to capture and store valuable data on the hypervisor, hosts and guest performance metrics (at the deepest level of granularity available), logs, tasks and events, hierarchy and topology, time data and more
- The ability to easily query, report and analyze data from multiple disparate systems to
  a. Quickly evaluate and categorize data that give you value vs. those that don't
  b. Navigate through all your data, using the topology familiar to system administrators
  c. Easily investigate problem areas, identify abnormalities and drill down into raw data and effortlessly reduce troubleshooting times, across the technology stack
  d. Visualize metrics and environment status in real time for immediate visibility into your virtual stack
  e. Provide results in context of both the physical and virtual components of your infrastructure, applications and services and connect them easily to gain end-to-end visibility
- Ad hoc trending and reporting for troubleshooting, capacity planning, resource trending analysis and optimization, complex correlation between configuration changes and overall operational intelligence

Big Data Analytics for Virtualization Using Splunk

There is a direct connection between big data and virtualization and a treasure trove of critical insights that can be gained via analyzing big data. The challenge lies in the collection, storage and analytics on these very large datasets with a single solution. Through monitoring, reporting and analyzing both real-time and historical machine-generated big data across all tiers of the virtualized datacenter, Splunk software provides business-critical insights that enable operational intelligence.

Whether it is located on-premise or in the cloud (public or private), Splunk Enterprise indexes machine-generated data, just as Google indexes the Internet, and does not use a database as a backend datastore, thus eliminating the need for pre-normalization or parsers. A schema-on-demand approach allows you to apply structure from data in the index when it's needed, at query time, versus relying on a rigid and brittle ETL/schema-based approach. The Splunk software approach allows for rapid search through terabytes of data and the Splunk Search Processing Language (SPL) allows you to map, visualize and correlate your most important data assets. The SPL supports five different types of correlation and over one hundred statistical commands.

Apps extend the core functionality of Splunk Enterprise. Over three hundred Apps and Inputs are available on the Splunkbase community site. Apps extend Splunk capabilities across technology silos (Cisco, Citrix, VMware, NetApp, NetScaler, WebSphere, etc.) and across use cases (Application Management, Security, Compliance, Capacity Planning, Transaction, Database Monitoring, etc.).

The Splunk App for VMware is the extension to Splunk Enterprise. The Splunk App for VMware focuses on collecting, storing and analyzing information from hypervisors and vCenter Server (VC). With the Splunk core ability to harness data from any technology layer and the Splunk App for VMware focusing on the virtualized VMware environment, the complete solution helps to find causal links and analyze and correlate information across the virtualization stack and connect the dots end-to-end.

Other similar apps for virtualization such as Splunk for Server Virtualization (which provides visibility into XenServer and HyperV) and Splunk for Desktop Virtualization (Splunk for Citrix XenDesktop and Splunk for Citrix XenApp) make it easier to analyze, visualize and trend data for operational intelligence across multiple technologies.

With the Splunk apps for virtualization, you can:

Collect and store in extensive granularity: Most existing administrative tools for virtualization aggregate performance information into five minute summaries after a few hours. For instance, the Splunk App for VMware collects granular information (such as performance metrics) at 20-second intervals from each ESX/I host and stores it in full fidelity within Splunk. The App collects and indexes logs, tasks, events, inventory and topology, enabling you to capture and store data from your VMware environments at appropriate levels of detail for as long as you want it. Splunk also provides a means to quickly sift and filter out data that you deem is unnecessary for your environment.

Search and query: Comprehensive data about your virtualized environment provides broad ranging benefits: troubleshooting, monitoring and alerting, capacity planning and optimization, business planning – game changing operational intelligence. Splunk provides an easy interface to search through the data quickly, allowing you to focus on data that is important and relevant. With your data in Splunk, you can easily identify which Virtual Machines (VMs) are getting a lion's share of the resources or if there is a large chunk of unallocated or unused resources going to waste. Use Splunk to identify access bottlenecks and make sure you are able to proactively alert any impending emergencies with both your physical and virtual environments. Drill down into raw data for diagnosis or trend over time to determine resource utilizations and contentions.

Analyze and Report: With out-of-the-box reports for critical insights into your virtualization solution, Splunk provides an intuitive interface that lets you easily customize existing reports and create new ones on the fly. Gain real-time insights, find historical trends and correlate data across both the physical and virtual layers for end-to-end reporting, planning, chargeback and analytics. Insight-oriented analytics from this information provides critical awareness across your virtual stack, allowing you to use your existing data more effectively.

Conclusion

With the broad adoption of virtualization for mission critical infrastructure, desktops and applications, and an increasing trend of enterprises superimposing multiple virtualization solutions from multiple vendors, a new class of solutions is needed to provide deeper monitoring and analytics that span the entire physical and virtual stack. A big data solution that focuses on collecting and consolidating data from various technology silos and providing a rich interface for advanced analytics is the most likely path to gaining operational and business insights.

Splunk helps organizations unlock the hidden value of their machine data regardless of the heterogeneity of the IT infrastructure. The proven ability of Splunk software to collect, index and harness machine-generated data from physical IT environments and virtualized infrastructures allows organizations to search and investigate, monitor and optimize their infrastructure resources. It's time to apply big data analytics to the virtualized datacenter to enable deeper insights for IT and the business. Discover the depth and value of your IT assets across your enterprise and gain this unprecedented visibility and operational insights using Splunk.

Free Download

Download Splunk for free. You'll get a Splunk Enterprise license for 60 days and you can index up to 500 megabytes of data per day. After 60 days, or anytime before then, you can convert to a perpetual Free license or purchase an Enterprise license by contacting sales@splunk.com.

1 blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
2 http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/
3 http://blogs.forrester.com/mike_gualtieri/12-05-17-whats_your_big_data_score

250 Brannan St, San Francisco, CA, 94107 info@splunk.com | sales@splunk.com 866-438-7758 | 415-848-8400 www.splunkbase.com